A Critical Perspective On Microarray Breast Cancer Gene Expression Profiling

Herman M.J. Sontrop

Herman Michaël Johannes Sontrop

A Critical Perspective On Microarray Breast Cancer Gene Expression Profiling

Proefschrift

ter verkrijging van de graad van doctor
aan de Technische Universiteit Delft
op gezag van de Rector Magnificus prof. ir. K.C.A.M. Luyben
voorzitter van het College voor Promoties
in het openbaar te verdedigen op donderdag 15 januari 2015 om 12:30 uur
door

Herman Michaël Johannes SONTROP

doctorandus in de econometrie
geboren te Nijmegen

Technische Universiteit Delft

Dit proefschrift is goedgekeurd door de promotor:
Prof. dr. ir. M.J.T. Reinders,

Copromotor:
dr. ir. P.D. Moerland

Samenstelling promotiecommissie:

Rector Magnificus,	voorzitter
Prof. dr. ir. M.J.T. Reinders,	Technische Universiteit Delft, promotor
Dr. ir. P.D. Moerland,	Academisch Medisch Centrum, Amsterdam, copromotor
Prof. dr. L.F.A. Wessels,	Nederlands Kanker Instituut
Prof. dr. T. Heskes,	Radboud Universiteit Nijmegen
Prof. dr. C.J.F. ter Braak,	Plant Research International
Dr. M.J. Jonker,	Universiteit van Amsterdam
Prof. dr. G.W. Klau,	Centrum voor Wiskunde en Informatica; Vrij Universiteit Amsterdam
Prof. dr. ir. G. Jongbloed,	Technische Universiteit Delft, reservelid

Dr. ir. W.F.J. Verhaegh heeft als begeleider in belangrijke mate aan de totstandkoming van het proefschrift bijgedragen.

The work in this dissertation was supported by Philips Research Laboratories and carried out at the High Tech Campus, Eindhoven; Philips Research at Briarcliff New York and the Academic Medical Center, Amsterdam.

Cover design: Proefschriftmaken.nl —— Uitgeverij BOXPress
Printed & Lay Out by: Proefschriftmaken.nl —— Uitgeverij BOXPress
Published by: Uitgeverij BOXPress, 's-Hertogenbosch

Cover image: illustration of the double helix structure in a deoxyribonucleic acid (DNA) molecule. DNA is a molecule that encodes the genetic instructions used in all known living organisms. Image downloaded from http://mashable.com/category/dna.

ISBN: 978-94-6295-075-7

CONTENTS

CHAPTER 1

INTRODUCTION

1.1 Breast cancer

Cancer represents a large collection of diseases that causes cells in the body to change and eventually grow out of control (Weinberg, 2007). In most instances, over time the cells form a mass called a *tumor*. Tumors may be divided into two broad categories, i.e. tumors that have the ability to grow into other tissues or spread to distant parts of the body and those that do not. These are respectively called *malignant* and *benign* tumors. Even though benign tumors can be problematic, e.g. they can become very large or press on healthy organs, they are mostly not life threatening. Therefore, benign tumors are generally not considered to be cancerous. Cancers are often named after the site where the tumor originates. Major sites include the lung, colorectum, prostate and breast. The primary form of cancer considered in this dissertation is *breast cancer*, i.e. the formation of malignant tumors that originate from the breast. Breast cancers represent the most common form of cancer in women worldwide and are the leading cause of death for women in the western world (Siegel *et al.*, 2011). Although breast cancer is observed in men as well, it is more than one hundred times more common in women (http://www.cancer.org/). A major concern in the context of cancer is *metastasis*, i.e. the ability of a tumor to spread out to other sites of the body and to form new tumors that replace normal tissue. Unfortunately, at present the prognosis of metastatic breast cancer is still very poor and forms the actual cause of death in approximately 90% of the cases (www.breastcancerdeadline2020.org). Figure 1.1 presents a schematic overview of the problem of metastatic breast cancer. An important source of information for cancer research are the levels of gene expression that can be inferred from removed tumor tissue (Figure 1.2). For a long time researchers did not have the proper tools to study the expression of genes on a large scale. This, however, all changed with the advent of the microarray.

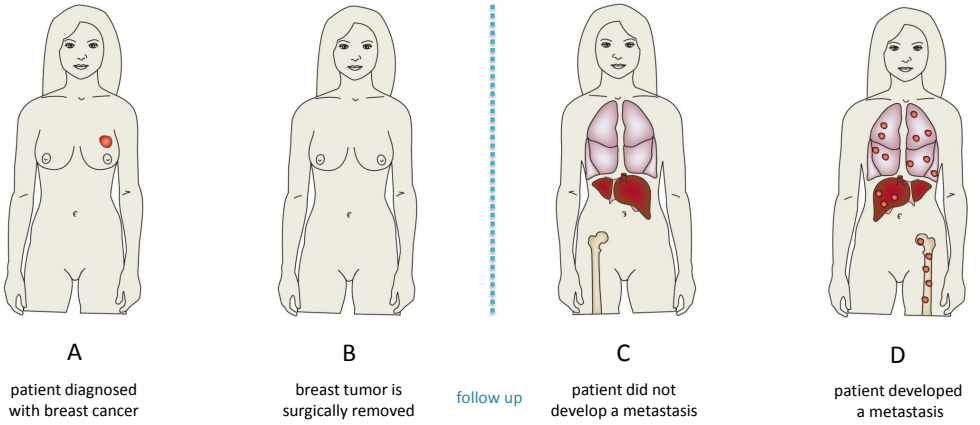


Figure 1.1: The problem of metastatic breast cancer. Panels A and B depict the scenarios in which a patient is diagnosed with breast cancer and after which the primary tumor is removed, respectively. Prior to the removal of the tumor, a patient may receive neo-adjuvant breast cancer therapy. In approximately 70% of the cases, the tumor does not metastasize (van't Veer *et al.*, 2002), (Panel C). However, in approximately 30% of the cases, despite the removal of the primary tumor, the patient develops a metastasis (Panel D). Unfortunately, in most instances this leads to the death of the patient. From a pathological perspective, however, tumors with markedly different clinical outcomes often look the same. An intriguing question therefore is to what extent it is possible to extract information from the removed tumor tissue, in order to obtain an improved perspective on the prognosis of the disease. Figure adapted from Sotiriou and Piccart (2007).

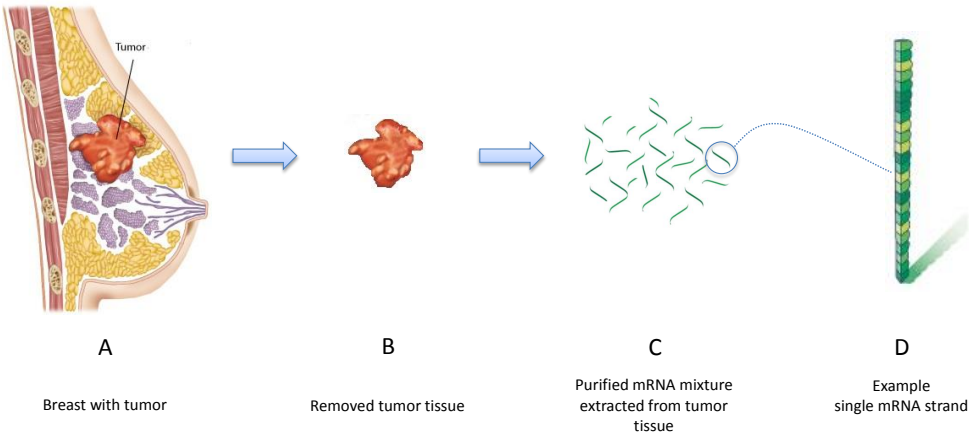


Figure 1.2: The goal of a breast cancer gene expression profiling experiment is to quantify how much each gene is expressed, based on the corresponding amount of mRNA in an mRNA mixture extracted from removed tumor tissue. Panels A-C depict the extraction process, while Panel D shows a toy example of a single mRNA strand of interest. Image Panel A taken from http://www.patientresource.com .

1.2 Microarray technology

A crucial development in modern research into cancer and other diseases was the *human genome project*. The project started around 1990 and set out to determine the sequence of the approximately three billion base pairs that make up human DNA, as well as to identify all its genes (Watson, 1990). The project was largely completed in 2003, with over 99% of the human genome decoded (Collins *et al.*, 2003). A surprising realization was that the human genome seemed to contain far less genes than previously conjectured, i.e. only 20 to 30 thousand, instead of the expected one hundred thousand. The sequence knowledge obtained by the human genome project made it possible to develop exciting new measurement techniques with unprecedented detection power. One such technique, introduced in the mid-nineties, is the *microarray* (Schena *et al.*, 1995; Lockhart *et al.*, 1996). In contrast to more traditional methods, which could only measure the expression levels of a few genes in parallel, microarrays allowed for the simultaneous assessment of the expression levels of many thousands of genes within the same experiment.

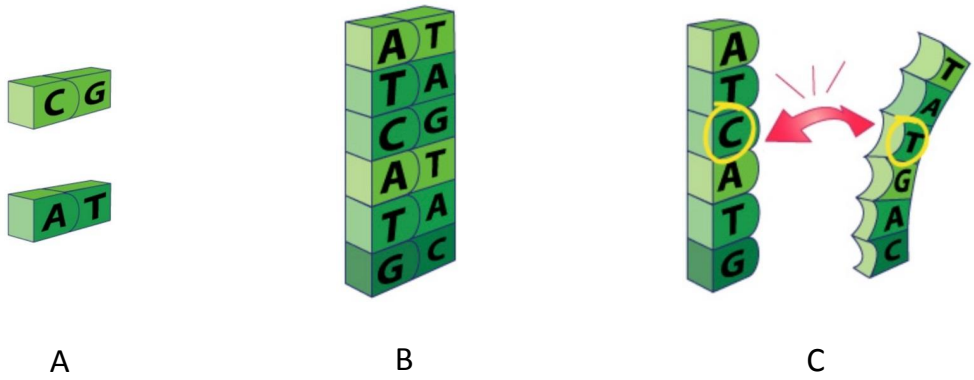


Figure 1.3: Concept of complementary base pairing. Panel (A): Only four molecules, i.e. *bases*, make up every DNA chain: adenine [A], guanine [G], thymine [T] and cytosine [C]. Complementary bases bind together when in proximity. C pairs with G and A pairs with T. This fundamental type of attraction is the working horse of the microarray. Panel (B): Example of two complementary strands that bind together, i.e. *hybridize*. Panel (C): When there is a mismatch, however, the possibility to hybridize is strongly reduced. Image taken from http://www.affymetrix.com.

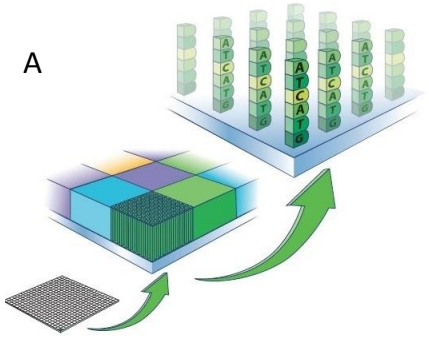
Microarrays heavily rely on a concept known as *complementary base pairing* or *hybridization* (Figure 1.3). They exploit the fact that two DNA strands that are each others complement bind together to form a double-stranded molecule. In a microarray experiment, one of the two strands is immobilized on a solid state substrate. We refer to the immobilized strand as a *probe*, which typically represents a specific gene coding region. The complementary strand, which,

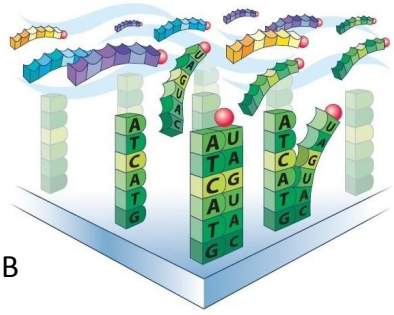
for the moment, we assume to be in the extracted mRNA mixture, is referred to as the *target*.

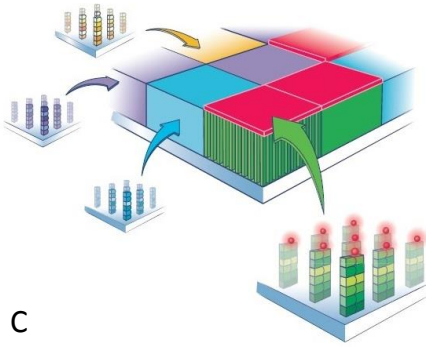
When the target in the RNA mixture comes in proximity to its immobilized counterpart, hybridization occurs. In practice, millions of identical probes are placed on the array, which all target the same gene. These, however, only occupy a small amount of the available surface. By filling the surface systematically with probes from many distinct genes, a genome-wide perspective on gene expression can be obtained (Figure 1.4)[1]. The key idea is that the more RNA fragments for a certain gene are present, the more hybridization occurs and therefore a higher intensity reading is obtained. Thus, the intensity readings represent a measure of the abundance of a transcript.
Various forms of microarrays exist, e.g. single channel and dual channel, spotted and oligonucleotide based arrays etc., which all come with their own protocols, image processing and pre-processing techniques. An in-depth discussion on the differences between the various types of microarrays available is beyond the scope of this introduction. For more information on microarrays in general

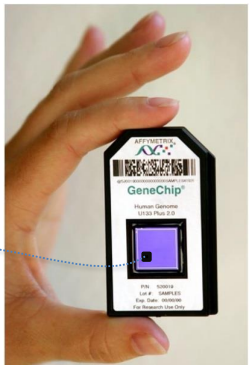
[1]Most steps in Figure 1.4 are specific to Affymetrix GeneChips. Of note, most probes on Affymetrix GeneChips come in two forms, i.e. *perfect match* (PM) probes and *mismatch* (MM) probes. PM probes have a sequence exactly complementary to their target gene. MM probes differ from the perfect match probe by a single base substitution at the center base position. MM probes were thought to be representative for non-specific hybridization. However, their added value is questionable and many researchers prefer to ignore these probes entirely (Irizarry *et al.*, 2003). For additional information on the use of MM probes see Bolstad (2004); Gautier *et al.* (2004) or Gohlmann and Talloen (2010).

Figure 1.4 *(facing page)*: Conceptual view on the workings of a microarray. (A): Individual probes are systematically immobilized on the array surface into a large set of probe cells, shown as colored tiles. Within a probe cell there are millions of copies of the same oligonucleotide. (B): During hybridization, a cocktail of fluorescently labeled purified RNA fragments is washed over the array surface. Note that Affymetrix works with cRNA, therefore in the target strand thymine (T) has been replaced by uracil (U) c.f. Figure 1.3. (C): After the hybridization stage has been completed, the array undergoes a series of washing and staining steps. Subsequently, each probe cell is scanned by a laser. This excites the fluorescent staining agents, from which an intensity measurement is obtained, quantifying the level of hybridization at the probe location, i.e. probe cell. (D): Example of a real-life microarray, i.e. Affymetrix Human Genome U133A GeneChip. (E): Toy example of the image file obtained after scanning the array. Each probe cell in panel C is represented by 64 pixels in the image file. Of these the border pixels are discarded. For each probe cell, a probe intensity is computed by taking the 75th percentile of the remaining 36 center pixel intensities. A gene may be represented by multiple probes, each targeting a different subsequence of the gene. For a single gene these are depicted as red dots. To overcome possible spatial defects in the array, such probes are spread out across the array. (F): In the final gene expression matrix, for each array j, the measurements related to all probes for gene g are collapsed into a single value. Images A-D taken from http://www.affymetrix.com.

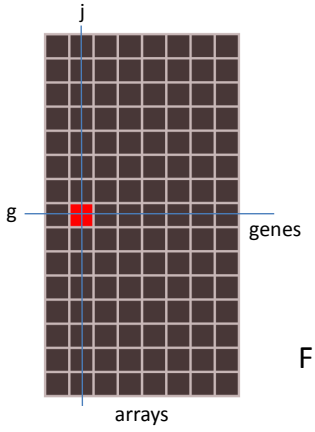

A


B


C


GeneChip
D


E


j
g
genes
arrays
F

we refer the reader to Speed (2003) or Causton *et al.* (2009). Most of the data analyses in this dissertation, however, are based on data from a single type of microarray, i.e. the *Affymetrix GeneChip* (Figure 1.4 D). For more information on these types of arrays see, Affymetrix (2002), Gautier *et al.* (2004), Irizarry *et al.* (2006).

1.3 Issues with microarrays and their analysis

Given their enormous potential, microarrays were quickly adopted by the research community. However, as microarrays became more popular, it became apparent that the accurate analysis and interpretation of microarray data provided a plethora of unique challenges (Simon *et al.*, 2003). The following subsections describe some of the many challenges in working with microarray measurements and factors that influence these measurements.

1.3.1 Complexity of obtaining measurements

From a biological, as well as a technical perspective microarray technology is complex. Various steps are needed to go from an extracted tumour tissue to a set of usable microarray-based gene expression measurements (Figure 1.5). These steps can roughly be broken up into two parts, i.e. an *in vitro* part (steps A-M), and an *in silico* part (steps O-R). Image analysis (step N) is at the interface of the in vitro and in silico parts. In the first part one directly works with biological material. In the last part one no longer works with biological material, but with intensity measurements extracted from the image file obtained from scanning the array. In this dissertation we exclusively look at steps related to the in silico part. Problematic issues during in vitro steps, therefore, have to be dealt with in downstream data analyses. Finally, note that even though Figure 1.5 already contains a substantial number of steps, the statistical analysis of the data yet has to start.

1.3.2 Influence of time, RNA degradation and sample storage

Time plays an important role in many microarray studies. Consider the tumour extraction step (Figure 1.5, step A). Regardless of the measurement technique, as a tumour develops over time, the gene expression measurements depend on the time of diagnosis and moment of tumour removal. From a technical perspective we may distinguish between *warm ischemic duration*, i.e. the time between surgical incision and tumour specimen removal and *cold ischemic duration*, i.e. the time from tumour specimen removal to sample preservation. The influence of these steps on expression profiling has been studied by Hatzis

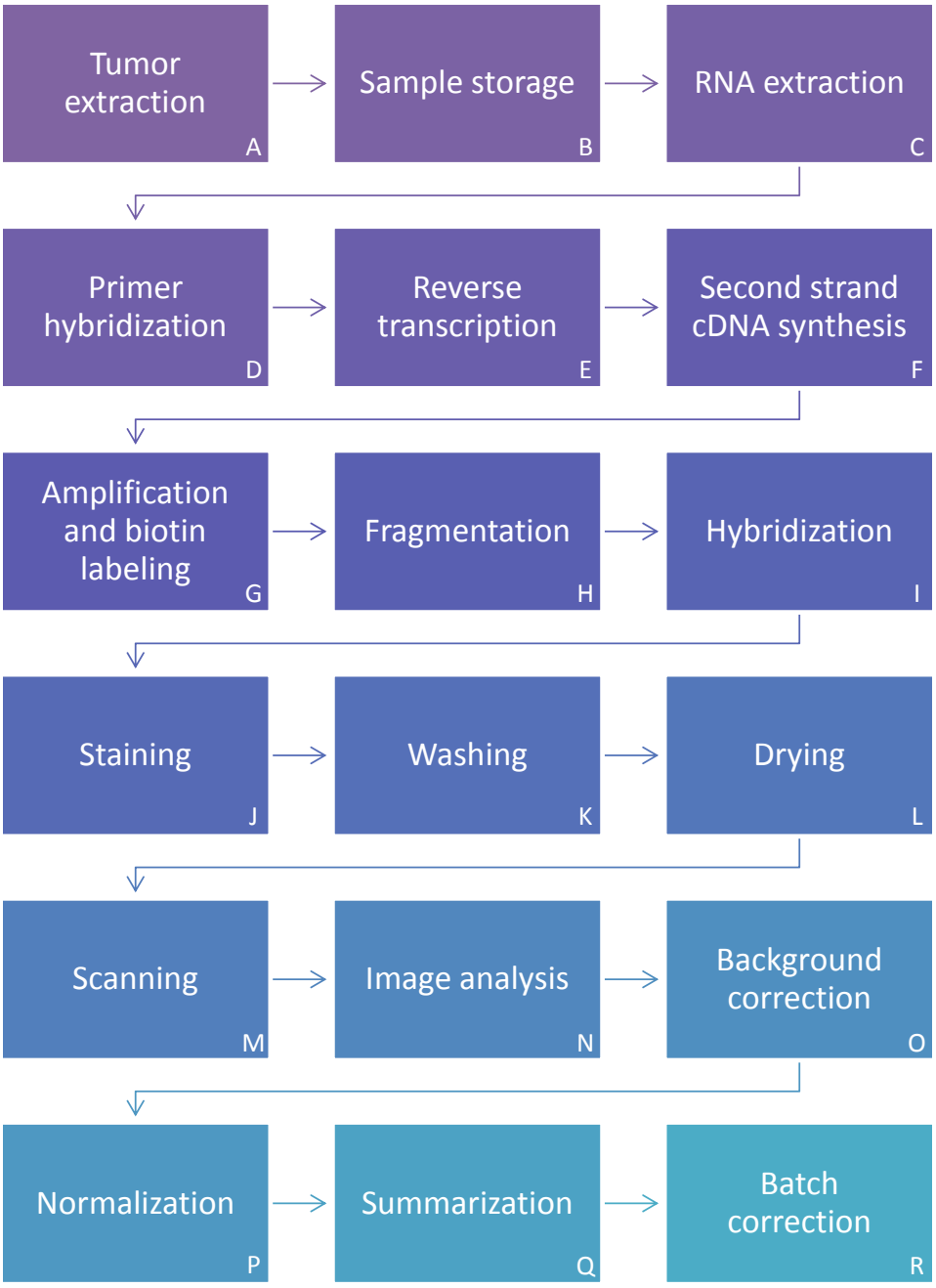


Figure 1.5: Condensed overview of the steps needed to go from an extracted tumour tissue to a set of usable microarray-based gene expression measurements suitable for statistical analysis. Individual steps are discussed in the running text. Furthermore, quality control (QC) can be considered at various steps and is therefore not indicated as a single step (see Section 1.3.5).

et al. (2011). The same authors also note that gene expression measurements may be influenced by the activation of cellular stress responses induced by surgical manipulation. With time the experimental conditions themselves may also change. For instance, during an experiment the atmospheric ozone levels may change, which in turn effects the microarray measurements (Fare *et al.*, 2003). A further complication is that over time RNA degrades. This has an impact on various steps in Figure 1.5. A more in-depth analysis on the implications of RNA degradation is offered by Copois *et al.* (2007). The way a sample is stored (Step B), may also influence the expression estimates. Medeiros *et al.* (2007) report that frozen tissues yield more intact RNA than formalin-fixed, paraffin-embedded tissues. Finally, the actual time a tumour has been in storage may negatively influence performance (Roepman *et al.*, 2005).

1.3.3 Wet lab steps

Steps D-R in Figure 1.5 depend strongly on the precise array design and platform, i.e. microarray technology used. Of these, steps D-N are wet-lab steps involving various detailed protocols and machines developed by the array manufacturers. A discussion of these is beyond the scope of this introduction. The precise execution, e.g. specific type of image scanner and its settings, however, has a considerable impact on the final expression estimates. The wet lab stage ends with scanning the array (step M) and performing primary image analysis (step N), e.g. gridding (Figure 1.6), and estimation of foreground and background intensity. Tu *et al.* (2002) performed a set of replicate array experiments in an attempt to separate the influence of steps related to sample preparation from steps related to the hybridization process, including the staining and scanning of the arrays. The authors found sample preparation to have only a modest impact. Hybridization noise, however, had a much stronger impact. Furthermore, hybridization noise was found to be dependent on the expression level itself, with different characteristics in the low and high end.

1.3.4 Pre-processing and batch correction

All actions related to the in vitro part (Figure 1.5, steps A-M) of a microarray experiment influence the intensity distributions observed at step N. These, however, reflect not only wanted biological variation, but also a substantial amount of unwanted non-biological variation. A major task in microarray data analyses thus is to effectively dissociate actual gene expression information, i.e. *biological variation*, from experimental noise and bias, i.e. *technical variation*. Fortunately, by borrowing strength over genes and arrays, a substantial part

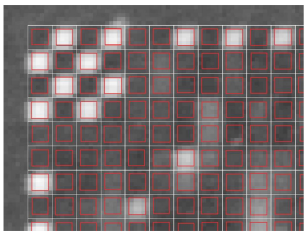


Figure 1.6: Part of an actual image file obtained from scanning an Affymetrix GeneChip. The checkerboard pattern in the left top corner is the result of the hybridization of various control probes, which are used to superimpose and align a grid on the image file. The grid is used during image analysis in order to locate the probes. Image segment obtained from Bolstad (2004).

of the unwanted systematic technical variation can be modelled and adjusted for.

During *pre-processing* (steps O-Q), one models and adjusts the data for unwanted technical variation. Popular pre-processing schemes for Affymetrix GeneChips are RMA (Irizarry *et al.*, 2003), MAS5.0 (Affymetrix, 2002) and dChip (Li and Wong, 2001). Even in the absence of a transcript hybridizing to a probe, a scanner will pick up low levels of fluorescence on the chip (Gautier *et al.*, 2004). During *background correction* (step O) we correct the data for such background noise. Furthermore, this step also corrects for non-specific binding, i.e. cross-hybridization (Bolstad, 2004). For Affymetrix GeneChips, the latter effect may also be handled by using information from mismatch (MM) probes. It is, however, controversial if MM probes provide reliable information (Bolstad, 2004). More advanced methods like GCRMA explicitly take probe sequence information into account to correct for non-specific binding (Wu *et al.*, 2004). During *normalization* (step P), we adjust the overall distributions of probe intensity values over arrays such that they become directly comparable. Differences in these distributions typically arise due to differences in the quantity of RNA extracted between tumours or systematic differences between chips (Gautier *et al.*, 2004). Pre-processing ends with a *summarization* step (step Q). During this step, one summarizes the values of all probes that target the same transcript into a single value.

Batch effects (Figure 1.5, step R) represent biases in the expression data caused by differences in processing groups and/or processing times (Leek *et al.*, 2010). These frequently arise when we combine data from different studies (and possibly different platforms) into a *compendium* (van Vliet *et al.*, 2008). Batch effects, however, may also be present in data from the same study, for example, due to day-to-day variation (Leek *et al.*, 2010). Batch effects

often cannot be removed by normalization alone (McCall and Irizarry, 2011). Various methods have been proposed for this step, see Johnson *et al.* (2007); McCall *et al.* (2010) or Teschendorff *et al.* (2011). We note that even though many effective adjustment schemes have been developed, ultimately, they all require making assumptions on the data (Scherer, 2009). These may or may not be met in practice and are often hard to validate, mainly due to a lack of ground truth data. Therefore, the validity of the final expression data after all adjustments have been made, remains a concern.

1.3.5 Quality and reproducibility of measurements

Microarrays are delicate devices. The array surface, for instance, can easily be damaged by scratches or dust (Scherer, 2009). Due to the complexity of the microarray construction process, it is not uncommon that the array itself is faulty. Unfortunately, prior to hybridization it is hard to tell whether an array is of insufficient quality. By using information from fitted probe-level models, in combination with information on the physical probe locations, a reasonable assessment can be made on the quality of the hybridization (Bolstad, 2004). Figure 1.7 presents various examples of poor quality hybridizations that can be identified by such schemes. To a certain degree effects as seen in these figures can be corrected for during summarization by down-weighting probes in problematic regions. Over time, many additional quality measures have been developed, see Kauffmann and Huber (2010); McCall *et al.* (2011). Fully automated quality control, however, remains challenging.

The reproducibility of microarray measurements has been a long standing issue. The issue was actively studied by the MicroArray Quality Control (MAQC) project, which claimed good inter- and intraplatform reproducibility of gene expression measurements (Shi *et al.*, 2006, 2010). The authors, however, note that data quality is also largely driven by the skill and technical knowledge of the experimenter. The MAQC study itself, however, has been criticized for basing its conclusions on too few and overly clean reference samples (Liang, 2007). Another way to gain a perspective on the issue of reproducibility, is by performing a *self-self hybridization*. In these, the same mRNA is used in two separate hybridizations. Note that in this context we do *not* expect any differential expression. The top panel of Figure 1.8 shows an example of a self-self hybridization. From the figure we see that that microarray measurements can be rather crude, i.e. two-fold changes between replicate arrays are, unfortunately, not uncommon.

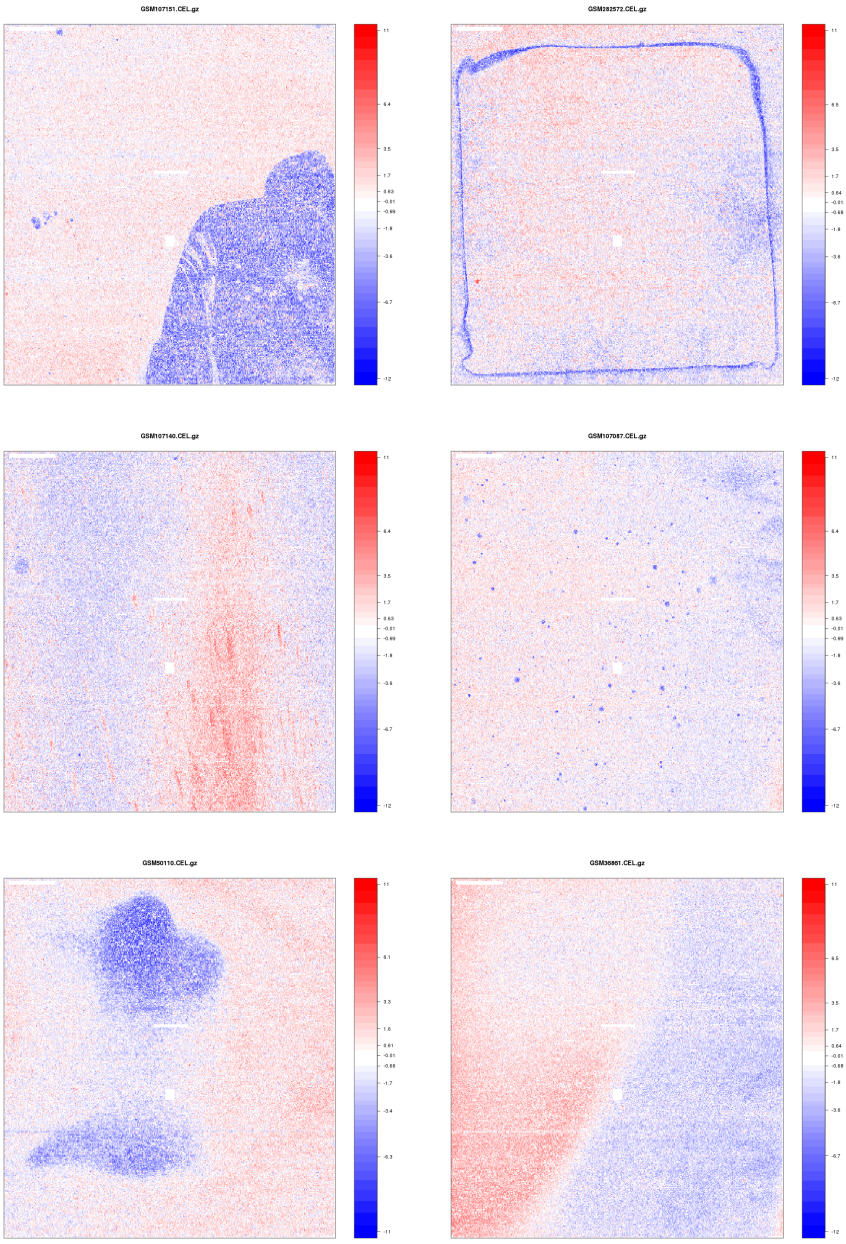


Figure 1.7: Examples of poor quality hybridizations. The figure shows a variety of *chip pseudo images* (Bolstad *et al.*, 2005). In these, for each array separately, the residuals of a fitted probe-level model are plotted at x and y coordinates which correspond to the physical locations of the probes on the array. Here positive and negative residuals are depicted in red and blue colors, respectively. Due to the way probes are laid out on the array surface, we do not expect to see a clear spatial structure in the residuals. The images, however, clearly show various forms of spatial artifacts. These could have been caused by, e.g. dust, scratches, bubbles in the hybridization chamber or, for instance, by performing hybridizations on an uneven surface.

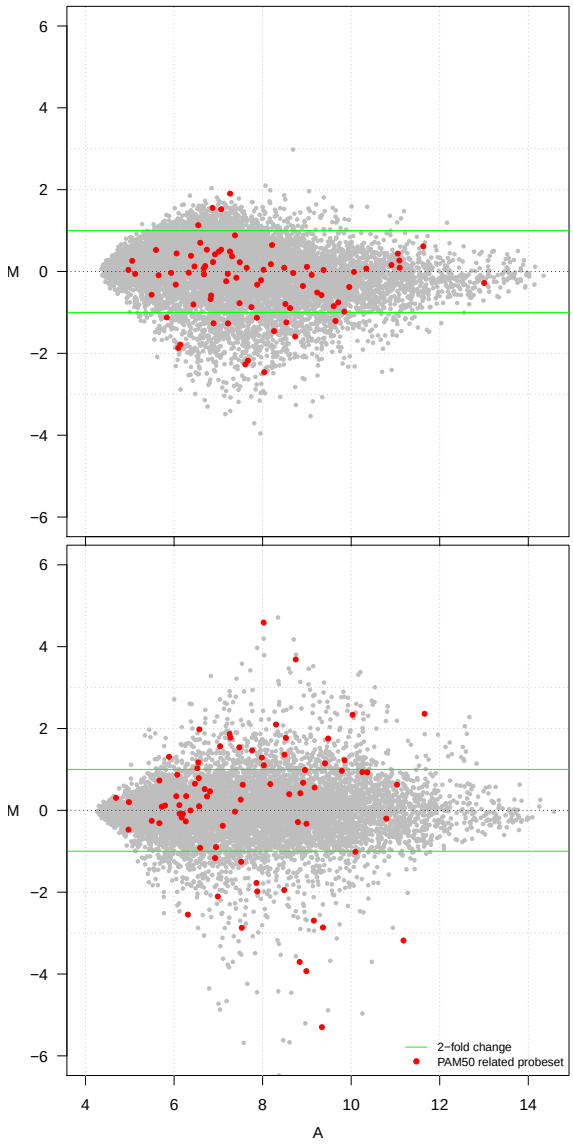


Figure 1.8: Example on measurement noise. The top panel shows an MA-plot (Dudoit *et al.*, 2002b), of the gene expression measurements of 10,000 randomly selected probes from two replicate arrays, based on Affymetrix GeneChips, normalized by frozen RMA (McCall *et al.*, 2010). M: difference between the log 2 expression values, A: average of the log 2 expression values. The red dots represent all probesets related to the 50 genes in the PAM50 subtype predictor, proposed to distinguish between different breast cancer subtypes (Parker *et al.*, 2009) (see Chapter 3). The horizontal green lines at -1 and 1 indicate an absolute fold-change of 2. Note that for technical replicates we ideally expect all probes to have an M value of zero. Many probes, however, show an absolute fold change larger than two, including various probes in the PAM50 gene list. For comparison, the bottom panel shows an MA-plot related to expression data from two distinct patients. Array (GEO) accession numbers - top panel: gsm36858.cel.gz, gsm308396.cel.gz, bottom panel: gsm38054.cel.gz, gsm519118.cel.gz.

1.4 Microarrays and breast cancer

The specific combination of performing breast cancer research with microarray technology was put firmly on the map by two landmark papers, i.e. *Molecular portraits of human breast tumours* by Perou *et al.* (2000) and *Gene expression profiling predicts clinical outcome of breast cancer* by van't Veer *et al.* (2002). Even though both papers involve the analysis of genome-wide gene expression levels based on microarray technology, their goals and methodologies are quite different. In machine learning terms these boil down to performing unsupervised and supervised learning, respectively. The goal of *unsupervised* learning is to describe the associations and patterns among a set of input measurements. In *supervised learning*, the goal is to predict the value of an outcome variable, i.e. a class label, based on a number of input measurements (Friedman *et al.*, 2001). A key difference between the two techniques is that the outcome variable is typically *not* available in unsupervised learning.

1.4.1 Perou et al.

Perou *et al.* (2000) mainly focussed on the application of *unsupervised learning* in order to divide, i.e. cluster, breast cancer samples of similar pathology into various *breast cancer subtypes*. Note that a priori the subtypes are unknown. In statistical terms they may be thought of as latent variables. In total, 65 surgical specimens of human breast tumours from 42 different individuals were analysed, using microarrays representing 8,102 human genes. Twenty of the tumours were sampled twice, once before and once after treatment. By hierarchical clustering of the microarray gene expression data, four main subtypes were distinguished, i.e. luminal, basal-like, Human Epidermal Growth Factor Receptor 2 positive (HER2+) and normal breast tumours. Striking differences in observed expression levels were reported between samples from different subtypes. These were primarily driven by estrogen receptor (ER) status, which strongly separated luminal samples (ER+) from HER2+ and basal-like samples (ER-). Furthermore, Perou et al. note that gene expression patterns in two tumour samples from the same individual were almost always more similar to each other than either was to any other sample.

1.4.2 Van 't Veer et al.

Van 't Veer et al. (2002) argue that breast cancer prognosis can already be derived from a gene expression profile obtained from the primary tumour. The authors observe that only 30% of all patients who are diagnosed with breast cancer, ultimately develop a metastasis. Thus for many patients additional treatment can potentially be avoided. Classical clinical indicators such as node

Figure 1.9: Main breast cancer subtypes considered in this dissertation, i.e. the *intrinsic molecular subtypes* (Parker *et al.*, 2009).

status or histological grade, however, fail to accurately classify breast tumours according to their clinical behaviour. A cohort was analyzed of 78 patients who were all previously diagnosed with breast cancer. After the removal of the primary tumour, 34 of the patients developed a metastasis within 5 years, while 44 did not. Note that here, a priori, the outcome measure, i.e. the set of class labels, is known. DNA microarray analysis was performed in a supervised way and resulted in a 70-gene signature, strongly predictive of metastasis. In a 19 sample validation cohort, 5 out of 7 good prognosis and 12 out of 12 poor prognosis cases were correctly predicted.

1.4.3 Intrinsic subtypes

Over time the subtyping taxonomy introduced by Perou *et al.* (2000) has been further refined, see Sørlie *et al.* (2001, 2003); Hu *et al.* (2006); Parker *et al.* (2009). An important finding in these works is that the luminal subtype can be refined into at least two subtypes, i.e. luminal A and luminal B. The resulting five subtypes are often referred to as the *intrinsic molecular subtypes* (Parker *et al.*, 2009) (Figure 1.9). The subtypes mainly differ in their ER, HER2, progesterone (PGR) and proliferation status. Compared to other subtypes, the luminal A subtype is associated with a relatively good prognosis and is mainly characterized by high expression of ER-regulated genes and low proliferation (Wirapati *et al.*, 2008; Prat *et al.*, 2011). The other subtypes mostly represent highly proliferating tumours. The main scheme used in breast cancer literature to identify intrinsic subtypes is by means of a *single sample predictor* (SSP, Figure 1.10). Over the years various SSPs have been developed, i.e. the SSPs of Sørlie *et al.* (2003), Hu *et al.* (2006) and more recently, Parker *et al.* (2009). The latter SSP is also known as PAM50. These SSPs, as well as a variety of other subtype predictors, are extensively addressed in Chapter 3 of this dissertation.

1.5 Statistical challenges

Even though the papers by Perou *et al.* (2000) and van't Veer *et al.* (2002) had an enormous impact on the breast cancer research community, over time, various authors have expressed their concerns with respect to the findings reported in these works. As it turns out, the challenges in working with microarray data do not end once we have obtained a set of reliably normalized measurements. A major hurdle in working with microarray data is that the number of samples (n) is typically much lower than the number of available features (p). The number of samples is often in the range of tens to hundreds, while the number of features is in the order of tens of thousands. The low sample size causes various statistical problems, in both supervised and unsupervised learning schemes. These mainly involve the problems of proper model fitting and validation, and lack of stability. Over the years it has become clear that especially the combination of these aspects is problematic.

1.5.1 Challenges and solutions in unsupervised learning

Various studies (Tibshirani and Walther, 2005; Pusztai *et al.*, 2006; Haibe-Kains *et al.*, 2012), have reported severe stability problems in fitting hierarchical clustering models as suggested in the landmark paper by Perou *et al.* (2000). Lusa *et al.* (2007) report on the difficulties of using SSPs on external datasets. Note that in order to use an SSP, the expression values in the external cohort must be brought to the same scale as the SSP centroid data. This is often achieved by gene centering. However, the distribution of ER+ and ER- cases in the external dataset itself may have a strong impact on the centering step and may bias subsequent subtype assignments. Furthermore, the assignment of samples to the intrinsic subtypes was found to be highly dependent on the selected SSP (Weigelt *et al.*, 2010a; Mackay *et al.*, 2011). Based on these findings, Weigelt *et al.* (2010a) conclude that SSPs do not reliably assign samples to subtypes and are therefore not ready for a clinical implementation. An interesting alternative to SSP-based subtyping was proposed by Desmedt *et al.* (2008) in the form of Subtype Classification Models (SCMs). SCMs rely on *model-based clustering* of module scores. The *module scores* reflect the activity of important biological processes such as ER and HER2 signaling. Compared to hierarchical clustering based approaches, SCMs show strongly improved stability properties (Haibe-Kains *et al.*, 2012). SCMs are extensively discussed and analysed in Chapters 2 and 3 of this dissertation.

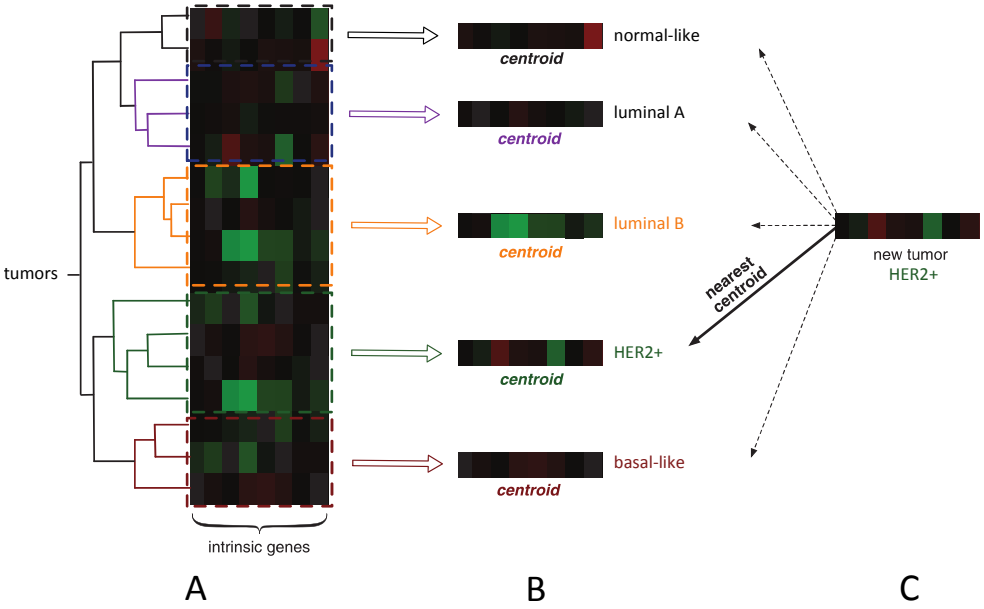


Figure 1.10: Single sample prediction. (A): on a training cohort unsupervised learning is performed to identify subtypes by hierarchical clustering, using only a subset of the genes. In breast cancer literature, this gene list is referred to as an *intrinsic gene list* (IGL) (Hu *et al.*, 2006). The resulting dendrogram is cut such that we obtain the number of desired subtypes, here five. The dashed colored lines indicate the identified clusters. (B): subsequently, for each subtype a *centroid* is computed. A centroid contains for each gene in the IGL the average expression value for all samples of the same subtype. (C): new cases are predicted to be of the subtype of their nearest centroid, i.e. an SSP is a *nearest centroid predictor*. In the example, the closest centroid is the HER2+ centroid. Therefore, we predict the new case to be of subtype HER2+. Figure adapted from Haibe-Kains *et al.* (2012).

1.5.2 Challenges and solutions in supervised learning

The main problem in supervised learning is that in a $p >> n$ scenario, classic predictor construction schemes cannot be used out of the box, since they require $p < n$ (Hastie and Tibshirani, 2004). Two frequently adopted solutions in this scenario are to either lower the dimension of the feature space, e.g. by applying a univariate filter step, or to apply some form of regularization during model fitting (Guo *et al.*, 2007). Another issue in small sample settings is the lack of validation data, which makes model fitting prone to overfitting (Mitchell, 1997). We speak of *overfitting* when a model performs poorly on validation data, despite excellent performance on training data. A frequently applied strategy to combat overfitting is the use of cross-validation (Friedman *et al.*, 2001). However, the use of cross-validation in combination with feature selection may be problematic. Simon *et al.* (2003) report a strong bias in performance estimation when we first use class label information of *all* available

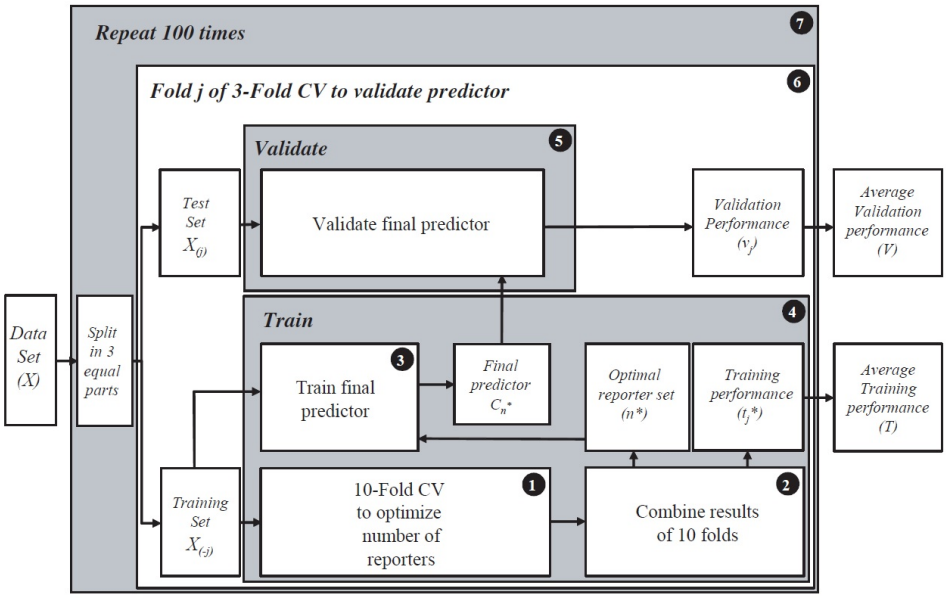


Figure 1.11: Unbiased performance estimation requires complex protocols. Schematic representation of the double-loop cross-validation protocol by Wessels *et al.* (2005). An extension of this protocol in the context of subtyping is offered in Chapter 4.

samples to identify a set of suitable features, and then construct predictors out of these in a cross-validated way. This phenomenon is known as *selection bias* (Ambroise and McLachlan, 2002). In this scenario even on random data very low error rates can be obtained. An exposition on the influence of selection bias on the Van 't Veer data is provided in Simon *et al.* (2003); Zhu *et al.* (2008)[2].

Many early studies reported overly optimistic results, largely due to inadequate validation. Unfortunately, these include the study by van't Veer *et al.* (2002). Michiels *et al.* (2005) re-analysed data from seven high-profile cancer outcome prediction studies, using a multiple random validation strategy. The authors found that, despite the high performance reported in the original papers, five out of the seven investigated studies did not classify patients better than chance. In addition, the gene lists identified were highly unstable and strongly depended on the selection of patients in the training sets. Similar findings were reported by Ein-Dor *et al.* (2005) on the Van 't Veer cohort. To thoroughly evaluate predictors, Wessels *et al.* (2005) propose a double-loop cross-validation protocol (Figure 1.11). The protocol uses an inner cross-validation loop to train

[2]The results of Van 't Veer on the external validation set are not biased due to methodological errors. However, the sample size of their validation cohort was too small and not representative for the general population of breast cancer patients.

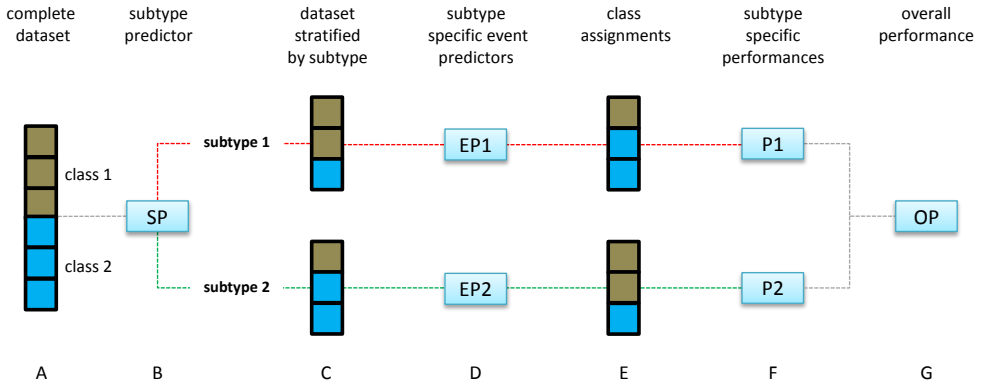


Figure 1.12: Subtype-specific event prediction and evaluation (A): dataset comprised of two classes, e.g. metastasis within five years or not. Colors indicate the class distribution. (B): subtype predictor (SP). Here we assume there are only two subtypes. (C): the application of the subtype predictor leads to a dichotomization of the samples. The top and bottom halve indicate samples predicted to be of subtype 1 (dashed red line) and subtype 2 (dashed green line), respectively. (D): for each subtype separately, an event predictor (EP) is obtained. (E): the application of the event predictors leads to two sets of class assignments. (F): for each subtype separately, a performance estimate is constructed (P). (G): the assignment information over the subtypes is translated into an overall performance estimate (OP).

a predictor, i.e. to determine the signature size, composition and to estimate any model parameters, while an outer cross-validation loop is used to assess predictor performance. In this work the Van 't Veer cohort was associated with a substantially lower performance than initially reported. Performance can be improved by increasing the sample size of the training cohort (Michiels *et al.*, 2005; van Vliet *et al.*, 2008), however, again not to the level reported in van't Veer *et al.* (2002)[3].

1.6 Contributions of this dissertation

In this dissertation we systematically study microarray-based event prediction, subtyping and combinations of these. As we have seen, many aspects, be it of technical, biological or statistical nature, may influence the outcome of a microarray experiment. We provide a number of carefully devised protocols,

[3]The results in van't Veer *et al.* (2002) suggest a sensitivity of $12/12 = 1$ and a specificity of $5/7 = 0.714$. This translates into a balanced accuracy rate (*bar*), i.e. the mean of the sensitivity and specificity, of 0.885. Wessels *et al.* (2005), however, only report a *bar* of 0.627. On pooled training data van Vliet *et al.* (2008) report a *bar* of 0.652. Of note, the results in Wessels *et al.* (2005) and van Vliet *et al.* (2008) are based on an extended sample cohort, including data from van de Vijver *et al.* (2002). They all, however, target the same sample population and therefore we may expect similar performance estimates.

by which the influence of important sources of variation can be isolated and explicitly quantified, even in the absence of a gold standard by which performance can be measured. Instead of applying these protocols to data from small spike-in or dilution studies, we applied them to a large collection of real-life breast cancer datasets of considerable size. The analyses in this work focus on three topics i.e. *(i)* the influence of feature variability on microarray breast cancer event prediction, *(ii)* the estimation of breast cancer subtypes and *(iii)* subtype-specific predictor construction and performance evaluation.

All analyses in this dissertation begin after the *in vitro* part of a microarray breast cancer experiment and primary image analysis (Figure 1.5, steps A-N) have been completed. A crucial next step in any microarray study is the pre-processing (Figure 1.5, steps O-Q) of the resulting intensity data. During pre-processing one attempts to correct for unwanted technical variation introduced during the in vitro steps. Various schemes have been developed for this task. These, however, make different assumptions on the data and rely on different modeling techniques. In the first part of this dissertation we study the impact of pre-processing on event prediction and feature selection, an important aspect of predictor construction[4].

Over time various studies have reported on the association of breast cancer subtypes with survival (Sørlie *et al.*, 2003; Wang *et al.*, 2005; Weigelt *et al.*, 2010d). As noted in Section 1.4.3, especially the luminal A subtype is associated with a relatively good prognosis. Given the relation of these subtypes with survival, instead of targeting all patients with a single predictor, we may first stratify the patients by subtype and subsequently construct an event predictor for each subtype separately, as depicted in Figure 1.12. Prior to this work surprisingly few authors have attempted such schemes. One of the first subtype-specific event prediction schemes in this regard is the 76-gene signature proposed by Wang *et al.* (2005). The signature is comprised of two parts, i.e. a 60-gene signature, for ER+ cases, and a 16-gene signature, for ER- cases[5]. Even though the signature showed a strong performance on ER+ cases, the performance was weaker on ER- samples (Foekens *et al.*, 2006; Haibe-Kains, 2009). The prevalence of ER-, however, is also notably lower than that of ER+ breast cancer. Note that in general a stratification by subtype implies a potentially strong reduction in the number of samples available for predictor construction. As sample size matters (van Vliet *et al.*,

[4]Other in silico steps such as batch correction (Figure 1.5, step R), alternative normalization schemes and overall quality control are addressed at various points in this dissertation, notably in the discussion chapter.

[5]Note that in Wang *et al.* (2005) subtypes were simply characterized by ER status.

Figure 1.13: Major steps in breast cancer research related papers.

2008), it is therefore questionable whether the potential benefit of subtyping can outweigh the drawback of a severe loss in sample size induced by the stratification.

As a prerequisite for subtype-specific predictor construction one has to be able to determine a set of relevant breast cancer subtypes that can be estimated realiably. In the second part of this dissertation we therefore extensively study the ability to assign samples to relevant breast cancer subtypes. Subtype assignments were based on Single Sample Predictors (Figure 1.10), but also on more recent subtyping approaches. Even though we restricted ourselves to a set of commonly accepted intrinsic subtypes, i.e. basal, HER2, luminal A and luminal B, their identification is already quite troublesome.

In the last part of this dissertation we study the evaluation of subtype-specific event prediction, based on divide and conquer schemes as depicted in Figure 1.12, from both an experimental and a theoretical perspective. In order to properly evaluate the potential benefits of subtype-specific event prediction, we provide an extended version of the Wessels protocol (Figure 1.11), with a special emphasis on the aforementioned sample size implications due to subtype stratification. Our protocol allows for a sound comparison of subtype-specific predictors with a baseline predictor that does not utilize subtype information. Furthermore, it allows for a proper comparison of individual subtype-specific predictors. With respect to performance evaluation we show that a divide and conquer approach brings various new statistical challenges. For a variety of frequently encountered performance measures from machine learning we provide a decomposition of the overall performance into subtype-specific performances. These show that the relation between subtype-specific and overall performance can be highly complex and counterintuitive.

1.7 What this work is and isn't

Figure 1.13 provides an overview of the main steps taken in most breast cancer research papers. Most of these works focus heavily on the final step, i.e. the biological interpretation of their findings. In this dissertation we, however, do *not* focus on biological interpretation, nor do we investigate the added value of

microarray-based signatures to standard clinical variables. Instead, we strongly focus on the various technical limitations and issues of microarrays and the consequences of these in subsequent analyses. Furthermore, performance evaluation is mostly analysed from a machine learning perspective, instead of a clinical perspective.

1.8 Dissertation outline

Figure 1.14 presents a condensed overview of the main topics treated in the remaining chapters in this dissertation. **Chapter 2** provides a comprehensive sensitivity analysis of the influence of feature variability in microarray breast cancer experiments. We define *feature variability* as either *pre-processing variability*, i.e. variation in the value of a feature induced by a switch to an alternative pre-processing scheme or *perturbation variability*, i.e. the variation in the value of a feature as caused by adding noise, based on the uncertainty information associated with the expression data point estimates. On a breast cancer compendium of over 1,100 hybridizations of both one and two-color array technology, we studied the influence of feature variability on *feature selection* and *event prediction*, based on data from six state-of-the-art pre-processing methods. Our experiments show that signature composition is unstable and strongly influenced by feature variability, even if the array platform and the stratification of patient samples are identical. Furthermore, we show that there is often a high level of discordance between individual class assignments for signatures constructed on data coming from different pre-processing schemes, even if the actual signature composition is identical.

Chapter 3 provides an in-depth comparison between SSP and SCM-based assignments of the intrinsic subtypes, motivated by the different and sometimes conflicting views on SSP-based subtyping expressed in the literature. Based on a carefully devised breast cancer compendium comprising over 4,000 hybridizations, we provide a comprehensive re-assessment of the concordance of a variety of previously published SSP and SCM predictors. Furthermore, we performed an extensive analysis of subtype predictors that were specifically designed to be highly concordant on the individual sample level. These were constructed via a semi-supervised approach on a set of *consensus sets* (CS). The CSs represent samples which were concordantly subtyped across a variety of re-fitted SCMs, the PAM50 subtype predictor, and the St. Gallen subtype scheme (Goldhirsch *et al.*, 2011). Three key ingredients of predictor construction were studied, i.e. choice of CS, gene list and predictor type used. Our experiments show that both SSP and SCM-based subtyping can achieve almost perfect levels of agreement. However, they also reveal that differences in selected gene lists and predictor types, may result in subtype assignments

	application			analysis perspective	
	event prediction	subtyping	sensitivity analysis	experimental	theoretical
Chapter 2	✓		✓	✓	
Chapter 3		✓	✓	✓	
Chapter 4	✓	✓		✓	✓
Chapter 5	✓	✓			✓

Figure 1.14: Brief overview of the chapters in this dissertation.

which show only substantial levels of concordance. Surprisingly, the influence of a change in CS was less than a change to another gene list.

Chapter 4 provides a comparison of subtype-specific breast cancer event predictors and a baseline predictor that does not utilize subtype information. Factors like unequal class distributions, differences in the number of samples per subtype, and sample size in general, however, highly complicate comparisons between predictors. We present an extension of the Wessels protocol (Figure 1.11), which allows for a fair comparison of these predictors. In our protocol, differences in sample size, class- and subtype distributions are carefully controlled. The protocol was applied to a breast cancer compendium comprising over 1,500 individual cases, with SCM-based subtypes. Two types of cohorts were investigated, i.e. the complete (unbalanced) compendium and a set of balanced compendia. In the latter, the prevalence of each subtype and the negative-positive class ratio per subtype were forced to be equal. Our experiments show that subtype-specific predictors outperform those that do not take subtype information into account, especially when taking sample size considerations into account. The advantage of subtype-specific predictions was largest on the balanced cohorts. On the unbalanced cohort the advantage of subtype-specific event prediction was notably smaller.

In **Chapter 5** we study the relation between subtype-specific and overall performance from a more theoretical perspective. For a variety of frequently encountered performance measures from machine learning, we provide decompositions of the overall performance into subtype-specific performances. We show that for certain performance measures, e.g. accuracy, precision or recall, the overall performance is a simple linear combination of the individual subtype performances. For these measures, an improvement of the performance of any subtype-specific predictor implies an improvement in overall performance. However, for other performance measures like the balanced accuracy rate, area under the ROC curve or the concordance index, additional cross terms appear in the combination of the subtype performances. The cross terms heavily depend on both the overall class imbalance and the subtype class imbalances. For such performance measures, improving subtype performances may actually result in a decrease of the overall performance.

Chapter 6 offers a discussion of various elements of the research presented

in this dissertation. Furthermore, we discuss several other research directions that we explored which did not lead to conclusive results. In addition, several alternative high-throughput measurement modalities and techniques to gene expression microarrays are discussed. We conclude with a reflection on microarray breast cancer profiling.

CHAPTER 2

A COMPREHENSIVE SENSITIVITY ANALYSIS OF MICROARRAY BREAST CANCER CLASSIFICATION UNDER FEATURE VARIABILITY

2.1 Abstract

Background: Large discrepancies in signature composition and outcome concordance have been observed between different microarray breast cancer expression profiling studies. This is often ascribed to differences in array platform as well as biological variability. We conjecture that other reasons for the observed discrepancies are the measurement error associated with each feature and the choice of preprocessing method. Microarray data are known to be subject to technical variation and the confidence intervals around individual point estimates of expression levels can be wide. Furthermore, the estimated expression values also vary depending on the selected preprocessing scheme. In microarray breast cancer classification studies, however, these two forms of feature variability are almost always ignored and hence their exact role is unclear.

Results: We have performed a comprehensive sensitivity analysis of microarray breast cancer classification under the two types of feature variability mentioned above. We used data from six state of the art preprocessing methods, using a compendium consisting of eight different datasets, involving 1131 hybridizations, containing data from both one and two-color array technology. For a wide range of classifiers, we performed a joint study on performance, concordance and stability. In the stability analysis we explicitly tested classifiers for their noise tolerance by using perturbed expression profiles that are based on

uncertainty information directly related to the preprocessing methods. Our results indicate that signature composition is strongly influenced by feature variability, even if the array platform and the stratification of patient samples are identical. In addition, we show that there is often a high level of discordance between individual class assignments for signatures constructed on data coming from different preprocessing schemes, even if the actual signature composition is identical.

Conclusions: Feature variability can have a strong impact on breast cancer signature composition, as well as the classification of individual patient samples. We therefore strongly recommend that feature variability is considered in analyzing data from microarray breast cancer expression profiling experiments[1].

2.2 Background

Microarrays are a powerful tool for biologists as they enable the simultaneous measurement of the expression levels of thousands of genes per tissue sample (Amaratunga and Cabrera, 2004). One of the interesting applications of gene expression profiling is the identification of compact gene signatures for diagnostic or prognostic purposes, such as cancer classification. One of the first studies in this regard was the work of van't Veer *et al.* (2002), in which a prognostic 70-gene signature is identified, that can be used to assess whether a breast tumor is likely to metastasize or not. Signatures like the 70-gene signature of Van 't Veer are, in essence, comprised of two parts: a limited set of features and a classifier that maps a vector of feature values to a class label. Limiting the number of features has several advantages. For one, using too many features with flexible classifiers quickly leads to overfitted decision rules. The inclusion of irrelevant features can also substantially degrade the performance of some classifiers. Furthermore, understandability, efficiency, and cost also benefit from more compact rules.

Microarray breast cancer event prediction, however, has proven to be difficult, as few classification rules are able to obtain a balanced accuracy rate of over 70%, when properly validated (Wessels *et al.*, 2005; van Vliet *et al.*, 2008). These performance indicators are also often associated with wide confidence intervals (Michiels *et al.*, 2005). Furthermore, Ein-Dor *et al.* (2005) showed that signature composition strongly depends on the subset of patient samples used for feature selection. In recent years many different signatures have been

[1]This work was published as: HMJ Sontrop, PD Moerland, R van den Ham, MJ Reinders, WFJ Verhaegh (2009). A comprehensive sensitivity analysis of microarray breast cancer classification under feature variability. BMC Bioinformatics, 10:389.

proposed, mostly derived using different patient populations and/or array technologies. Although the overall performance of these signatures is comparable, there is often a high level of inconsistency between class assignments obtained using different signatures (Reyal *et al.*, 2008). This poses significant challenges for the use of gene expression classifiers in clinical routine. Although biological variability is conjectured to play a major role in the observed discrepancies, in this chapter we show that even in a very controlled setting, using identical arrays, patient samples, signature composition, and classifiers, still large discrepancies in performance and individual class assignments can be observed under two types of variability.

One of the challenging aspects of microarray data is that they are subject to various sources of technical variation, arising from the many experimental laboratory steps needed to get from a tissue sample to an array scan, such as array batch variability, dye incorporation, uneven hybridizations, probe-failure caused by dust or scratches, or washing conditions (Zakharkin *et al.*, 2005). Some noise factors bias large groups of measurements in a systematic way. Fortunately, most of this bias can be removed by proper preprocessing. Many preprocessing methods have been proposed to address these systematic biases.

The effectiveness of such procedures and the plausibility of their assumptions, however, depends on factors such as study design, the array technology being used, and the biological phenomenon under study (Kreil and Russell, 2005). Furthermore, even after correction for systematic effects by the preprocessing method, there remains a residual variance that is both array and feature specific and that can be substantial (Rattray *et al.*, 2006). Detailed error models have been proposed that attempt to quantify such uncertainty around the expression data point estimates, e.g. the Rosetta error model (Weng *et al.*, 2006). Such uncertainty information has been incorporated in differential gene expression analysis methods (Liu *et al.*, 2006), as well as in clustering analysis (Li and Wong, 2001), and principal component analysis (Sanguinetti *et al.*, 2005), often leading to more consistent results.

The impact of noise on the outcome of the statistical analysis of microarray data has been subject of debate. Tu *et al.* (2002) performed a detailed sensitivity analysis to separate noise caused by sample preparation from noise related to the hybridization process. The latter was identified to be the more dominant of the two. In addition, a strong dependence of hybridization noise on the expression level was reported. Based on data from the MAQC study (Shi *et al.*, 2006), however, Klebanov and Yakovlev (2007) claim that for Affymetrix arrays the magnitude of technical variation has been gravely exaggerated in the literature and that the effects on the results of statistical

inference from Affymetrix GeneChip microarray data are negligibly small. However, contradictory findings have been reported in Chen *et al.* (2007), based on the very same data. In addition, the MAQC study itself has been criticized for presenting their case in a best case scenario, using too few and overly clean reference samples (Liang, 2007). With regard to the impact of the choice of preprocessing method, it has been observed in differential expression studies that preprocessing can strongly influence whether a gene is detected to be differentially expressed or not (Hoffmann *et al.*, 2002; Irizarry *et al.*, 2006). Similar observations have been made for the influence of preprocessing on classification (Stafford and Brun, 2007; Verhaak *et al.*, 2006), albeit in a different and much smaller setting than the work presented here.

Although microarray data is known to be subject to the sources of variation described above, in microarray breast cancer classification studies the influence of the choice of preprocessing scheme and of the uncertainty around expression data point estimates are almost always ignored. In this chapter, we study the effect of these two types of variability of expression data on breast cancer classification in detail. We define *preprocessing variability* as the variation in the value of a feature as induced by switching to an alternative preprocessing scheme. *Perturbation variability* is defined as the variation in the value of a feature as caused by adding noise based on the uncertainty information associated with the expression data point estimates. Furthermore, *feature variability* is understood to be the variation in the value of a feature as caused by either preprocessing or perturbation variability.

This chapter presents a comprehensive sensitivity analysis of microarray breast cancer classification under feature variability. The experiments involve data from eight different studies, involving 1131 hybridizations, and containing data from both one and two-color array technology. We studied the impact of preprocessing and perturbation variability on feature selection, classification performance, and classification concordance for six different preprocessing methods. In addition, we performed a comprehensive stability analysis for a diverse set of classifiers, by explicitly testing these classifiers for their noise tolerance. Stability was quantified by the variation in class assignment of perturbed expression profiles, where the amount of perturbation is based on uncertainty information directly related to the selected preprocessing strategy. Our results indicate that even when using identical arrays and sample populations, preprocessing and perturbation variability have a strong impact on the classification of individual breast cancer samples, as well as on the composition of breast cancer signatures, especially when the number of features is low.

total	*labeled*	*good*	*poor*	*array platform*	*repository*	*accession*	*ref*
147	120	91	21	Affy 133A	GEO	GSE 7390	Desmedt *et al.* (2007)
96	62	41	21	Affy 133A	GEO	GSE 2603	Minn *et al.* (2005)
247	193	156	37	Affy 133A	GEO	GSE 3494	Miller *et al.* (2005)
156	142	120	22	Affy 133A	GEO	GSE 1456	Pawitan *et al.* (2005)
178	120	92	28	Affy 133A	GEO	GSE 6532	Loi *et al.* (2007)
123	86	63	23	Affy 133A	AE	E-TABM-158	Chin *et al.* (2006)
97	97	51	46	Agilent custom	-	-	van't Veer *et al.* (2002)
87	87	75	12	Agilent custom	-	-	van de Vijver *et al.* (2002)

Table 2.1: Overview of the eight datasets used. The column *total* contains the total number of hybridizations available, while the column *labeled* shows the number of samples that have a properly defined class label. The next two columns indicate the decomposition of this number into good and poor prognosis class memberships. The columns *repository* and *accession* list in what repository and under which accession number each dataset can be found.

2.3 Materials and Methods

2.3.1 Data

We performed our sensitivity analysis using a compendium of eight publicly available datasets, which all have been used to predict whether a breast tumor will metastasize within five years (poor prognosis) or not (good prognosis), based on gene expression data inferred from removed tumor tissue. In total these contain microarray data from 1131 hybridizations and for 907 samples class label information was available (Table 2.1). Some of the eight datasets initially had an overlap, either in patient samples or in hybridizations. The compendium of 907 arrays, however, contains no overlap, as all duplicate cases were removed. Data from the studies of Van 't Veer and Van de Vijver were obtained using two-color custom ink-jet oligonucleotide arrays produced by Agilent. Processed data for these datasets can be downloaded from http://www.rii.com/publications/2002/default.html. Like the original authors, we combined the two datasets. We refer to this combined dataset as the Rosetta dataset. The Rosetta dataset consists of 87 lymph-node negative samples of the Van de Vijver dataset and of the 78 training and 19 validation samples of the Van 't Veer dataset. Data for all other datasets was obtained using Affymetrix GeneChips and CEL files were downloaded from GEO (Barrett *et al.*, 2009) and ArrayExpress (Brazma *et al.*, 2003). A more comprehensive overview of the selected hybridizations is presented in the Supplementary Information corresponding to Sontrop *et al.* (2009). See also Additional File 1 online[2].

[2]See http://www.biomedcentral.com/1471-2105/10/389/additional for all online files corresponding to this chapter.

2.3.2 Preprocessing

For the Van 't Veer and Van de Vijver datasets, we used the publicly available expression estimates and corresponding error information based on the Rosetta error model (Weng *et al.*, 2006). In principle, the Rosetta error model is applicable to both one and two-color arrays. However, for this model no freely available implementation exists and hence for the Affymetrix datasets this model was not applied. For the datasets using Affymetrix GeneChips we generated expression data from the available CEL files based on five different, frequently used preprocessing strategies: MAS 5.0, mgMOS, its multi-chip version mmgMOS, RMA, and dChip. For preprocessing, all available hybridizations were used. This is especially relevant for the multi-chip models dChip, RMA, and mmgMOS, which benefit from having more arrays assuming all hybridizations are of similar quality. The dChip expression estimates are constructed using only the information of the PM-probes, which is the default choice for dChip. Affymetrix datasets were log-transformed and all probesets were median centered after preprocessing, for each dataset separately. The validity and benefits of this step are further discussed in van Vliet *et al.* (2008) and Kim (2009). Preprocessing for the Affymetrix datasets was performed in R (Ihaka and Gentleman, 1996) using Bioconductor (Gentleman *et al.*, 2004) packages affy (Gautier *et al.*, 2004) and puma (Pearson *et al.*, 2009). Table 2.2 provides a summary of the six preprocessing methods used.

2.3.3 Perturbation

After preprocessing, we get an expression estimate x_{ij} for each array i and each feature (gene) j. In fact, x_{ij} is usually stochastic, following some distribution D_{ij} with mean $\mu_{ij} = x_{ij}$ and standard deviation σ_{ij} reflecting the measurement uncertainty associated with the point estimate x_{ij}. We utilized the uncertainty information as captured by the distributions D_{ij} to generate perturbed expression profiles as alternatives for expression point estimates x_{ij}, in a similar fashion as presented in Li and Wong (2001). For each sample i, for each gene j in a given signature, we simply draw a new data point $\widetilde{x}_{ij}$ by sampling from the corresponding distribution D_{ij}. Complete perturbed training and validation sets can be constructed by repeating this process for all samples and genes.

The Rosetta model, mgMOS, and mmgMOS are specifically designed to provide a σ_{ij} that reflects the uncertainty of the complete preprocessing cascade. In these three models D_{ij} is a Gaussian distribution. For mgMOS and mmgMOS, the corresponding σ_{ij} values were obtained using the R package puma (Pearson *et al.*, 2009). For the Van 't Veer and Van de Vijver datasets, we used

method	*package*	*function*	$\log_2$	σ	*reference*
RMA	affy	expresso	yes	yes	Irizarry *et al.* (2003)
mgMOS	puma	justmgMOS	yes	yes	Milo *et al.* (2003)
mmgMOS	puma	justmmgMOS	yes	yes	Liu *et al.* (2005)
dChip	affy	expresso	no	yes	Li and Wong (2001)
MAS 5.0	affy	expresso	no	no	Affymetrix (2002)
Rosetta	-	-	-	-	Weng *et al.* (2006)

Table 2.2: Preprocessing overview. The column *package* indicates which R package was used to obtain the expression values, while the column *function* provides the name of the function used from the package. Column $\log_2$ indicates if the expression estimates as returned by the function are already on $\log_2$ scale or not. The column σ indicates if the function directly computes uncertainty information or not.

the published expression values. For the Van de Vijver data, the standard deviations σ_{ij}, as estimated by the Rosetta error model, were reported directly. For the Van 't Veer data σ_{ij} was not provided directly, but σ_{ij} could be recovered from the published p-value information (see Materials and Methods). MAS 5.0, dChip and RMA are not specifically designed to provide detailed error estimates, although some of the uncertainty associated with the point estimates can be derived from the summarization step in the preprocessing cascade. For RMA and dChip, the uncertainty corresponding to the summarization step can again be modeled by a Gaussian distribution. The estimated σ_{ij} values for these two models were obtained using the R package affy (Gautier *et al.*, 2004). For MAS 5.0 it turns out that the estimates follow a distribution closely related to a t-distribution. Although error information for MAS 5.0 is not available from affy directly, it can be computed from the information affy provides, see below.

2.3.4 Rosetta perturbation scheme

For the Rosetta data, the technical noise levels were estimated by applying the Rosetta error model (Weng *et al.*, 2006). This yields, for each sample i and each gene j, a log-ratio expression measurement that is normally distributed with mean x_{ij} and standard deviation σ_{ij}. For the data of Van de Vijver σ_{ij} was reported directly. However, for the data of Van 't Veer only x_{ij} and the p-value p_{ij} of observing a value for the log ratio more extreme than x_{ij} if the true log ratio is zero with measurement error σ_{ij} were given. In this case, we can compute σ_{ij} as

$$\sigma_{ij} = \frac{|x_{ij}|}{\mathrm{cdf}^{-1}(1 - \frac{p_{ij}}{2})}, \tag{2.1}$$

where cdf denotes the cumulative density function of a $N(0,1)$ distributed random variable. Equation (2.1) can be obtained by noting that for a normally distributed stochast X with mean 0 and standard deviation σ the two-sided p-value is given by

$$\begin{aligned} p &= 2 \cdot \Pr(X > |x|) \\ &= 2 \cdot (1 - \Pr(X \leq |x|)) \\ &= 2 \cdot (1 - \Pr(\frac{X}{\sigma} \leq \frac{|x|}{\sigma})) \\ &= 2 \cdot (1 - \text{cdf}(\frac{|x|}{\sigma})) \end{aligned}$$

Although Equation (2.1) provides the value of σ_{ij} for most pairs of x_{ij} and p_{ij}, in two cases we cannot deduce the value of σ_{ij} using (2.1). When p_{ij} equals 1 the denominator of (2.1) equals zero and hence the outcome is undefined. In that case, we conservatively set σ_{ij} to zero. Furthermore, we have to be careful of rounding effects if p_{ij} is close to one, since then the denominator of (2.1) is close to zero, which may yield unrealistically high values for σ_{ij}. To avoid this, we bound σ_{ij} from above by one. Since all reported expression values involved in the dataset of Van 't Veer are between -2 and 2, it makes sense to generate expression values in a similar range. A value of $\sigma_{\max} = 1$ ensures that with high probability we will generate a perturbed expression value between -2 and 2 for a gene with x_{ij} equal to zero. The same threshold was applied to the estimates corresponding to the study by Van de Vijver. Finally, for the Rosetta data, given the resulting standard deviations σ_{ij}, we perturbed the gene expression data by simply adding to each measurement x_{ij} some Gaussian noise with mean zero and standard deviation σ_{ij}:

$$\widetilde{x}_{ij} = x_{ij} + \epsilon_{ij} \qquad \text{with} \qquad \epsilon_{ij} \sim N(\mu = 0, \sigma = \min\{\sigma_{\max}, \sigma_{ij}\}). \tag{2.2}$$

2.3.5 MAS 5.0 perturbation scheme

In order to obtain perturbed expression values, we use the standard error of the probeset expression summary of MAS 5.0. The summarization step in MAS 5.0 works as follows (Affymetrix, 2002; Bolstad, 2004). Let the natural scale intensities of the different probes in a certain probeset and on a certain array be given by $\{y_1, y_2, \ldots, y_n\}$, with n denoting the number of probes in the probeset. Assume that these intensities are already background- and

PM-corrected via bg.correct and pmcorrect.mas in the affy package. In order to create a summary value for a probeset, MAS 5.0 calculates a robust average using the 1-step Tukey biweight estimator on $\log_2$ scale. The 1-step Tukey biweight estimator calculates a weighted average of the individual probes, using the bisquare function as a means to generate weights

$$w_k = \begin{cases} 0 & \text{if } |u_k| > 1 \\ (1 - u_k^2)^2 & \text{if } |u_k| \leq 1 \end{cases} \tag{2.3}$$

where u_k is defined as

$$u_k = \frac{\log_2(y_k) - M}{cS + \epsilon}. \tag{2.4}$$

with M the median of the $\log_2(y_k)$ values and S the median of the absolute deviations from M. Furthermore, c and ϵ are tuning constants, with default values of 5 and 0.0001, respectively. The probeset summary τ is obtained by creating the weighted average

$$\tau = \frac{\sum_{k=1}^{n} w_k \log_2(y_k)}{\sum_{k=1}^{n} w_k}. \tag{2.5}$$

using function tukey.biweight in expresso.

The MAS 5.0 preprocessing cascade ends with an intensity independent normalization step, in which each probeset summary on array i is multiplied by a scaling factor θ_i, such that the average intensity over all probeset summaries (on the natural scale) reaches a certain target intensity. The expression summary returned by MAS 5.0 for probeset j on array i then equals

$$2^{\tau_{ij}} \cdot \theta_i. \tag{2.6}$$

Similarly to one of the original studies in the compendium, Desmedt *et al.* (2007), we scaled all arrays to a target intensity of 600 using the function affy.scalevalue.exprSet.

We can compute the parametric based standard error σ_τ of τ as described in Affymetrix (2002); Bolstad (2004)

$$\sigma_\tau = \frac{\sqrt{\sum_{k=1,|u_k|\leq 1}^{n} (\log_2(y_k) - x)^2 (1 - u_k^2)^4}}{|\sum_{k=1,|u_k|\leq 1}^{n} (1 - u_k^2)(1 - 5u_k^2)|}. \tag{2.7}$$

In Affymetrix (2002); Bolstad (2004) it is suggested that perturbed τ values,

denoted by $\widetilde{\tau}$, can be obtained by

$$\widetilde{\tau} = \tau + rt(\nu) \cdot \sigma_\tau, \tag{2.8}$$

where $rt(\nu)$ represents a random draw from a t-distribution with ν degrees of freedom. The value for ν is set equal to $\max\{0.7(n-1), 1\}$; see Affymetrix (2002). Hence after $\log_2$ transformation and median centering, perturbed MAS 5.0 expression values $\widetilde{x}_{ij}$, for array i and probeset j, can be computed using Equations (2.6) and (2.8) as

$$\widetilde{x}_{ij} = \tau_{ij} + rt(\nu_{ij}) \cdot \sigma_{\tau_{ij}} + \log_2(\theta_i) - \text{med}_j. \tag{2.9}$$

where med_j represents the median expression level of gene j taken over all available hybridizations (see column *total* in Table 2.1).

2.3.6 dChip perturbation scheme

We used the PM-only version of dChip (Li and Wong, 2001). In this case, all estimated expression values are strictly positive and we can log-transform the data. Similarly to the default choice in dChip, we maintained a floor of one on the intensity data before log transformation. In rare cases where negative values arise after perturbing dChip-preprocessed data (last three columns of Table 2.3), we repaired the intensity value by setting it equal to one. Hence log-transformed median centered perturbed expression values $\widetilde{x}_{ij}$ for dChip expression estimates were computed as

$$\widetilde{x}_{ij} = \log_2(\max\{x_{ij} + \epsilon_{ij}, 1\}) - \text{med}_j \qquad \text{with} \qquad \epsilon_{ij} \sim N(0, \sigma_{ij}). \tag{2.10}$$

2.3.7 Stability measure: minority assignment percentage

Classification instability occurs when for a given classifier and a given sample, perturbed expression profiles are not all assigned to the same class. In order to quantify the instability over a large number of perturbed datasets, we propose the following simple stability measure, which we refer to as the *minority assignment percentage (map)* score. For a given sample and feature, we denote the percentage of perturbed datasets that lead to a classification into class 0 by p_0, and the percentage leading to a class label 1 by p_1. Then the minority assignment percentage is equal to $\min\{p_0, p_1\}$. In the ideal case, a map-score is equal to zero, indicating that all perturbed datasets lead to the same classification for this specific sample. In the worst case, it equals 50%, indicating that classification is purely random. Note that this observation is

author	*all*	all_1	all_2	*lab*	lab_1	lab_2	*fsel*	$fsel_1$	$fsel_2$
Desmedt	0	0.15	0.97	0	0.16	1.02	0	0.08	1.17
Minn	0	0.11	0.88	0	0.11	0.87	0	0.16	1.55
Miller	0	0.07	0.45	0	0.07	0.45	0	0.04	0.48
Pawitan	0	0.28	0.99	0	0.29	0.97	0	0.12	0.72
Loi	0	0.03	0.40	0	0.04	0.43	0	0.24	1.82
Chin	0	0.15	1.26	0	0.12	1.17	0	0.09	1.63

Table 2.3: Overview negative values in dChip. For each of the six datasets the column *all* gives the percentage of negative values, when taken over all probeset summaries and all available hybridizations. The columns all_1, and all_2 provide the same information as in *all*, on data for which we subtracted for each individual entry, one or two times the corresponding estimated standard error, respectively. The columns *lab*, lab_1, and lab_2 provide the same information as the previous three columns, when using only hybridizations that have a properly defined class label. Finally, columns *fsel*, $fsel_1$, and $fsel_2$ provide similar statistics, when using only the top-100 ranked features for each split, as obtained from the multi-rank strategy as described in the main text. The numbers represent averages over 50 splits.

independent of the choice of dataset, perturbation mechanism, classifier or number of features. In the remainder we will consider a classification to be unstable if the map-score exceeds a conservative threshold of 35%, meaning an almost random classification.

2.3.8 Sensitivity analysis protocol

All classification results are obtained in a systematic fashion, closely related to the protocol proposed in Wessels *et al.* (2005). Figure 2.1 provides a schematic overview of our workflow. Assume we have obtained expression values x_{ij} and the corresponding σ_{ij} values, for a given measure of expression, for some set of samples and a set of genes, using the methods described in the previous sections. In addition, assume we have selected an appropriate classifier, which we need to train. In the first step, we create a stratified split of the available data, in which 80% is used as a training set, while the remaining 20% serves as a validation set. In step 2, we create $P = 1000$ perturbed versions of the validation set. In step 3, we rank the features based on their Signal-to-Noise Ratio (SNR, see next section) on the (unperturbed) training set.
In the next step, we use the top-100 ranked features to construct a sequence of 100 classifiers, where the n^{th} classifier is constructed on the training data, using only the top-n ranked genes. At step 5, we invoke each classifier to obtain class assignments for both the unperturbed validation set, and for all perturbed versions. In step 6, we obtain a performance estimate for the unperturbed validation data by computing the balanced accuracy rate, that is, the average of the sensitivity and specificity. In step 7, we use the class

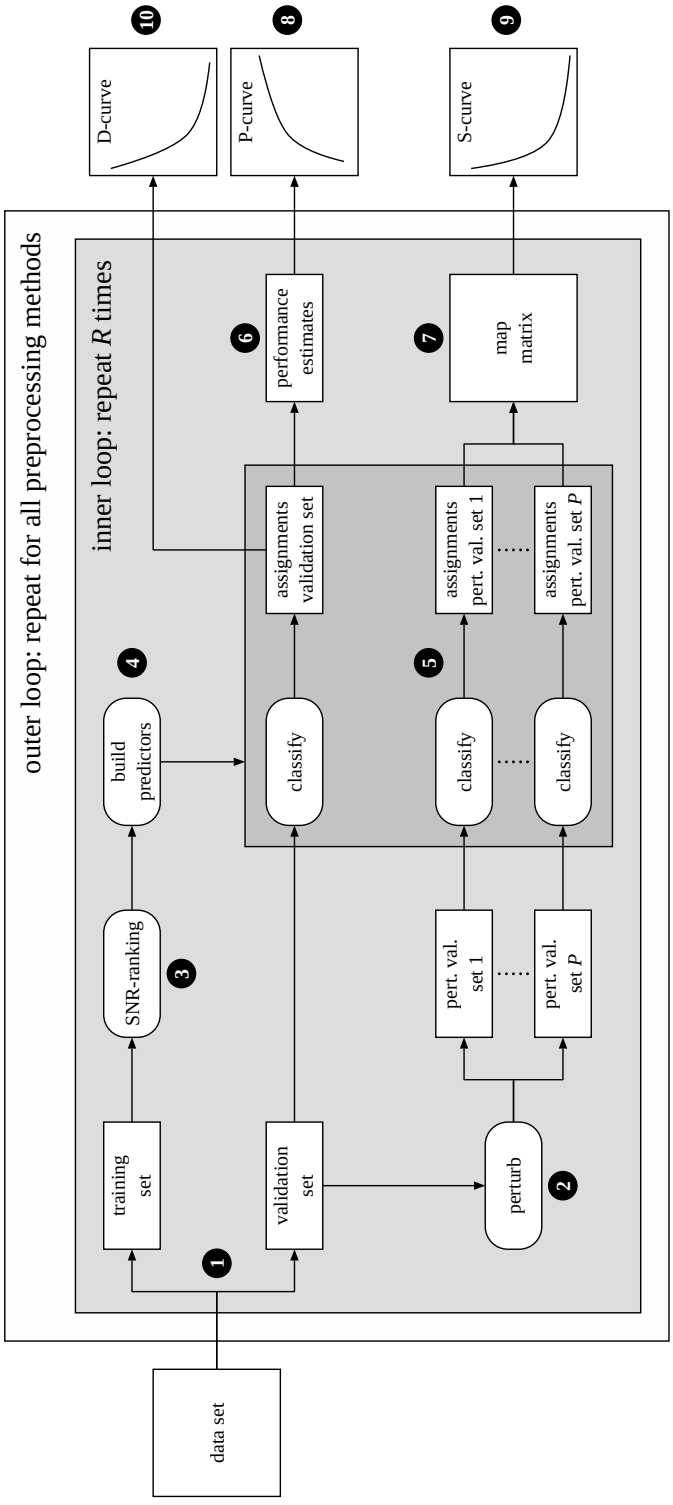


Figure 2.1: The sensitivity analysis protocol. For an explanation, see Section 2.3.8.

assignments of the 1000 perturbed validation sets, as obtained in step 5, to compute the associated map-scores and collect them in a *map-matrix*, where the entry at row i and column n represents the map-score of validation sample i, for a classifier trained on the top-n ranked features. To ensure that results are not split-specific, steps 1 to 7 are repeated $R = 50$ times (inner loop). At step 8, we compute a performance curve, referred to as the *P-curve*, which for each signature size $n \in \{1, \ldots, 100\}$, displays the average balanced accuracy over the R splits. Furthermore, as mentioned, for a given sample we consider a class assignment to be unstable if the corresponding map-score is larger than some threshold $T = 35$. For a given threshold, in step 9 we compute the stability curve, referred to as the *S-curve*, which for each signature size tells us the average percentage of cases, over R splits, that had a map-score larger than the selected threshold T. Note that ideally the S-curve should be zero for all entries. The whole procedure described above is repeated for each preprocessing method (outer loop). In order to compare results for different classifiers and preprocessing methods, for a given dataset and for each repeat of the inner loop we always used the same set of stratified splits. Finally, in step 10 we generate a discordance curve, referred to as the *D-curve*, for all distinct preprocessing method pairs. For a preprocessing method pair (m, m') and given classifier, the corresponding D-curve tells us for each signature size the average percentage of cases, over R splits, of inconsistent class assignments on the (unperturbed) validation sets. Similarly to the S-curve, ideally a D-curve is zero for all entries. Note that the map-scores used for the S-curves can also be viewed as a measure of concordance, under perturbation variability.

2.3.9 SNR-based feature rankings

As stated in the previous section, in the third step of our protocol we rank the available features based on their signal-to-noise ratios. For a given feature, let μ_0 and μ_1 denote the mean intensity value for class 0 and class 1, respectively, and let σ_0 and σ_1 be the corresponding standard deviations. Then the SNR is equal to

$$\text{SNR} = \frac{|\mu_0 - \mu_1|}{\sqrt{\sigma_0^2 + \sigma_1^2}} \tag{2.11}$$

Let $\text{SNR}_{j,m}$ denote the SNR value corresponding to gene j, based on data corresponding to preprocessing method m. In the construction of a signature we typically select the top-n features from such a ranking. Let $F_{n,m}$ denote the top-n genes, obtained using data from preprocessing method m, for a particular split.

Different preprocessing methods may lead to different lists of top-n genes. For two different methods m and m', a trivial measure to compare the lists $F_{n,m}$ and $F_{n,m'}$ would be to look at their intersection. From a classification standpoint, however, we would at least hope to obtain two lists that are of comparable strength. Let the total strength of a feature set F with respect to method m be defined as

$$S_m(F) = \sum_{j \in F} \mathrm{SNR}_{j,m}. \tag{2.12}$$

To compare two gene lists of cardinality n, we introduce the concept of *relative strength*, given by

$$\mathrm{RS}_n(m, m') = 100 \cdot \frac{S_m(F_{n,m'})}{S_m(F_{n,m})}. \tag{2.13}$$

The relative strength compares the total strength with respect to m for a selection based on m' to the selection based on preprocessing method m itself. As the latter gives the maximal total strength for a set of size n with respect to method m, the resulting relative strength will always be 100 at the most. Furthermore, since SNR values are non-negative, the relative strength is also non-negative. Note that a high relative strength implies that we expect a similar performance when using $F_{n,m'}$ as when using $F_{n,m}$. It does not imply that this performance is high per se.

2.3.10 Classifiers

In order to investigate whether the impact of variability is classifier specific, we employed a broad range of classifiers, being the nearest centroid (NC) classifier, k-Nearest Neighbors (k-NN) with $k \in \{1, 3\}$, a Support Vector Machine (SVM) with a linear kernel (SVMlin) and radial basis function kernel (SVMrbf), and the Random Forest (RF) classifier. For descriptions of the individual methods, see Duda *et al.* (2001); Breiman (2001). The NC and k-NN used a cosine based distance function (see Materials and Methods). All SVM results were obtained using the R package `e1071` and for each feature set a grid search was performed to find the best hyperparameter values. Classification results for RF were obtained using the R package `randomForest`.

2.3.11 Nearest mean classification using cosine distance

Let I^{g} denote the set of n_{g} samples belonging to the good prognosis class and I^{p} denote the set of n_{p} samples belonging to the poor prognosis class. The average good profile $\mathbf{m}^{\mathrm{g}}$ and the average poor profile $\mathbf{m}^{\mathrm{p}}$ are defined as

$$m_j^{\mathrm{g}} = \frac{1}{n_{\mathrm{g}}} \sum_{i \in I^{\mathrm{g}}} x_{ij} \qquad \text{and} \qquad m_j^{\mathrm{p}} = \frac{1}{n_{\mathrm{p}}} \sum_{i \in I^{\mathrm{p}}} x_{ij}.$$

A nearest mean classifier, using cosine correlation as its distance measure classifies a sample $\mathbf{x}$ to the good prognosis class if the distance of $\mathbf{x}$ to $\mathbf{m}^{\mathrm{g}}$, denoted by $d(\mathbf{x}, \mathbf{m}^{\mathrm{g}})$, is smaller than the distance of $\mathbf{x}$ to $\mathbf{m}^{\mathrm{p}}$, denoted by $d(\mathbf{x}, \mathbf{m}^{\mathrm{p}})$. From the definition of cosine distance it follows that a sample will be classified as having a good prognosis if and only if

$$\begin{aligned} d(\mathbf{x}, \mathbf{m}^{\mathrm{g}}) &< d(\mathbf{x}, \mathbf{m}^{\mathrm{p}}) \\ 1 - \frac{\mathbf{x}^T \mathbf{m}^{\mathrm{g}}}{\|\mathbf{x}\| \|\mathbf{m}^{\mathrm{g}}\|} &< 1 - \frac{\mathbf{x}^T \mathbf{m}^{\mathrm{p}}}{\|\mathbf{x}\| \|\mathbf{m}^{\mathrm{p}}\|} \\ \mathbf{x}^T \left(\frac{\mathbf{m}^{\mathrm{g}}}{\|\mathbf{m}^{\mathrm{g}}\|} - \frac{\mathbf{m}^{\mathrm{p}}}{\|\mathbf{m}^{\mathrm{p}}\|} \right) &> 0 \\ \mathbf{x}^T \mathbf{w} &> 0, \end{aligned} \tag{2.14}$$

where we define

$$\mathbf{w} = \frac{\mathbf{m}^{\mathrm{g}}}{\|\mathbf{m}^{\mathrm{g}}\|} - \frac{\mathbf{m}^{\mathrm{p}}}{\|\mathbf{m}^{\mathrm{p}}\|}.$$

Note that the classification rule (2.14) results in a linear classifier. The left-hand side of Equation (2.14) is usually referred to as the *discriminant score*.

Support vector machines

SVM results were obtained using the R package e1071. For each feature set, a grid search was performed to find the best parameters using the function tune. The parameter cost was varied from 0.5 to 5, with increments of 0.5. In addition, when using a radial basis kernel function, the parameter gamma equaled $\frac{2^x}{k}$, with x varying from -3 to 3, with increments of 1. Here k represents the number of features, which in our experiments varied between 1 and 100. Best parameters were selected using 3-fold cross validation, using tune.control. The parameter class.weights was used to set weights inversely proportional to the class frequencies, in order to compensate for the unbalanced class distributions.

Random forests

Random forest results were obtained using the R package randomForest. In randomForest the parameter ntree was set to 1000, while the parameters mtree and nodesize were set to their default values, the square root of the number of features and 1, respectively. During training the parameter sampsize was used in combination with the parameter strata to ward against unbalanced class distributions, by drawing an equal number of samples from the good and poor prognosis cases. In our case the number of samples to be drawn was always equal to the number of poor prognosis cases in the training set.

2.4 Results

The aim of our work is to get a comprehensive overview of the impact of feature variability on microarray breast cancer classification. We will operate under the null hypothesis that preprocessing and perturbation variability have no effect on feature selection and classification. Under this null hypothesis we expect that for different preprocessing methods or for perturbed versions of a dataset we 1) typically select the same features, 2) obtain identical class assignments and as a consequence 3) obtain overlapping P-curves and 4) obtain D-curves that are flat and close to zero. In addition, we expect to 5) obtain S-curves that are flat and close to zero as well. We first report our results of studying the impact of perturbation and preprocessing variability on feature selection, before moving on to their influence on classification.

2.4.1 Impact of feature variability on feature selection

In this chapter we focus on compact gene signatures. Unfortunately, feature selection on high-dimensional datasets, like the ones associated with microarray-based expression profiling, is typically unstable as different subsets of samples frequently lead to the identification of different feature sets (Ein-Dor *et al.*, 2005). From a classification perspective, such a difference does not necessarily signal a problem, as long as the performances of the sets are similar, although from a biological perspective it makes reasoning about the data much more challenging.

It has been observed that the impact of preprocessing strategies on differential expression detection is high (Hoffmann *et al.*, 2002). Note that feature selection strategies in microarray literature are often based on univariate ranking strategies, e.g. based on SNR-statistics or t-tests (Wessels *et al.*, 2005). One would expect that genes that are strongly differentially expressed are also highly ranked by univariate selection procedures and hence that feature selection is also influenced by feature variability. In this section we show several examples

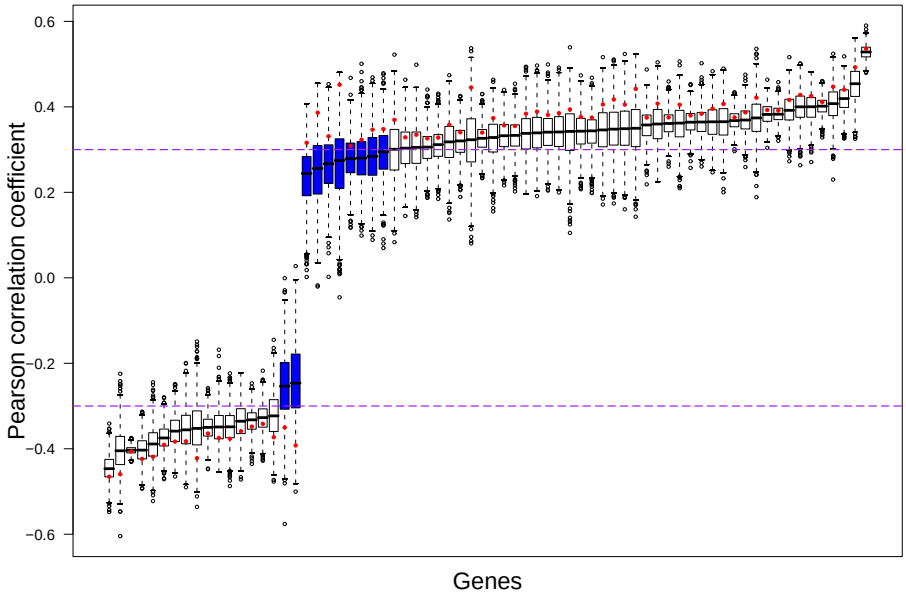


Figure 2.2: The impact of perturbation variability on the feature selection criterion for the 70 genes in the original signature of van't Veer *et al.* (2002). The dashed purple lines indicate the used absolute threshold of 0.3. Blue boxes indicate genes that do not meet this filter criterion in more than 50% of the perturbations. Results are shown over one 1000 perturbations. The red dots indicate the correlations obtained using the unperturbed expression values.

of the influence of perturbation and preprocessing variability on signature composition, i.e. feature selection. For the Rosetta data it was not possible to assess the influence of preprocessing variability, as for this dataset only processed data is publicly available.

Van 't Veer breast cancer signature composition is sensitive to perturbation variability

As a first example, consider the feature selection step used to identify the 70-gene breast cancer signature by van't Veer *et al.* (2002). This signature is comprised of the top-70 genes with an absolute Pearson correlation coefficient with the class label (0 or 1) larger than 0.3 as obtained from the 78 training samples of Van 't Veer. Note that the computation of correlation coefficients can be very sensitive to the presence of outliers. To test the sensitivity of this feature selection step, we created 1000 perturbed instances of the training set, using the Rosetta uncertainty estimates (see Material and Methods) and recomputed the Pearson correlation coefficients.
Figure 2.2 shows the sensitivity results of the feature selection step to perturb-

ation variability. We see that perturbation generally weakens the correlation of a gene with the class label vector. This is reflected by the red points, which were always located in the tails of the distributions. We also see that the correlations of weaker genes sometimes shrink to zero, indicating that they lose the connection with the class label vector. Although most genes will still be selected for most perturbations, there are ten genes, indicated by blue boxes, that would not have been selected for the majority of the perturbed training sets. Furthermore, the ranges of the correlation coefficients for the genes are quite large, implying that rankings based on them are unstable, as in Ein-Dor *et al.* (2005).

High impact of perturbation variability on feature rankings for Affymetrix datasets

In the previous example, the composition of the signature was given. In practice, however, the identification of a suitable set of marker genes is part of the discovery process. Our protocol, similarly to the protocol suggested in Wessels *et al.* (2005), employs a signal-to-noise ratio based ranking on each training split, in order to identify useful features for signature construction. This implies that the composition of our signatures is fully determined by the outcome of the ranking step and independent of the classifier used.
We examine the overlap between SNR-based rankings obtained using an unperturbed and a perturbed version of a dataset. Let $F_{n,m,k}$ denote the top-n ranked genes, using data from preprocessing method m, for split k and let $\widetilde{F}_{n,m,k}$ be the ranking obtained using a perturbed version. Although a complete overlap between these lists is preferable, we would at least hope to find a substantial part of the top half of one list in the other list. How large these parts are, is shown in Figure 2.3.
For most preprocessing methods the impact of perturbation noise appears to be large. Although the overlap increases when signature size increases, the overlap between a ranking based on unperturbed data and one based on perturbed expression data is generally less than 50%. In the Desmedt dataset there were two genes that almost always appeared on top of the SNR rankings in each split, which is the reason of the shape irregularity seen in the (blue) overlap curves for the study by Desmedt. For RMA, the overlap between rankings based on unperturbed and perturbed versions is much larger, with overlaps between 80 and 90%. In comparison to the other preprocessing methods RMA appears to give lower estimates on the measurement errors, although on the basis of our data one cannot tell if RMA underestimates the errors or if the other methods typically overestimate the errors.
Note that a lack in overlap does not necessarily signal a problem if the selected feature sets are of equal strength. Although the overlap for most preprocessing

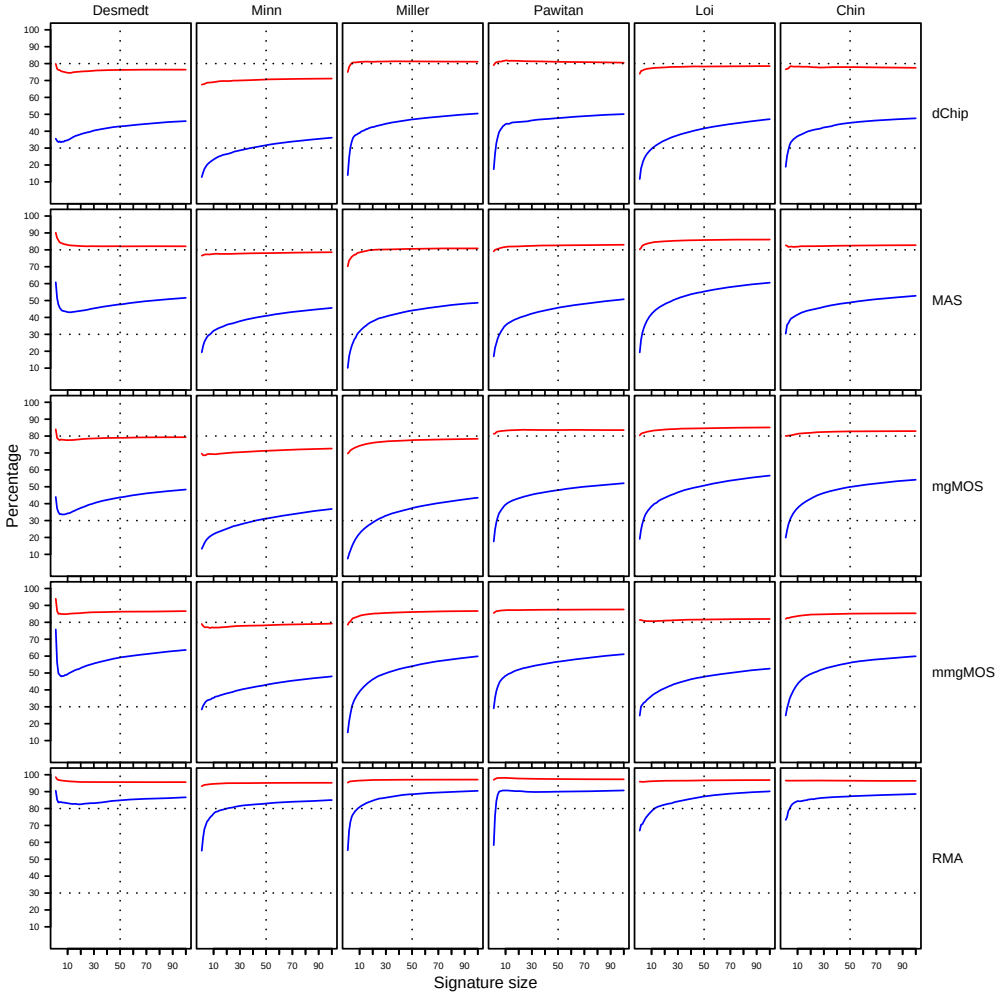


Figure 2.3: Impact of perturbation noise on gene rankings for the Affymetrix GeneChip data. Each dataset was split 50 times into a training and validation set, for which the validation set was subsequently discarded. Ranking was done only on the training sets. In addition, for each training set 50 perturbed versions were created and for each perturbation the overlap between $F_{n,m,k}$ and $\widetilde{F}_{2n,m,k}$ and between $\widetilde{F}_{n,m,k}$ and $F_{2n,m,k}$ was determined, yielding $50 \cdot 50 \cdot 2 = 5000$ overlap estimates for each list size n. The blue curves for each $n \in \{1, \cdots, 100\}$ provide the mean overlap taken over all corresponding estimates. The red curves indicate the associated average relative strengths between the feature sets $F_{n,m,k}$ and $\widetilde{F}_{n,m,k}$.

methods is quite low, the related relative strengths (see Methods section) are still high, with values of over 80% for most preprocessing schemes and values of over 95% for RMA, indicating that the performance for signatures based on the different rankings is expected to be comparable. A similar observation

was made in Ein-Dor *et al.* (2005), which for instance shows that on the Van 't Veer data the performance of the second best 70 genes was very comparable to the performance achieved by selecting the top-70 genes. Note that the latter two lists by construction have an overlap of zero. For some datasets many equally performing signatures exist, as was also noted in Roepman *et al.* (2006).

Affymetrix breast cancer signature composition is sensitive to preprocessing variability

Here we inspect the overlap between top-ranked feature lists, as obtained using different preprocessing methods i.e. we consider preprocessing variability. Consider two top-ranked feature lists, based on two different preprocessing schemes m and m', say of size 100, i.e. $F_{100,m}$ and $F_{100,m'}$. Similarly to the example in the previous section, we would hope to find a substantial part of the top half of one list in the other list. Figure 2.4A shows the overlap of the top-50 of one list in the top-100 of the other.

Different preprocessing strategies give rise to the selection of different features as well, as for all preprocessing pairs again none have a complete overlap. Within the same preprocessing family, i.e. mgMOS and mmgMOS, the overlap is high, although for the dataset of Loi there is already quite a discrepancy. For the remaining pairs we see that the overlap between top-ranked feature lists can be quite low. The overlap between different preprocessing families for the various datasets lies between 30 and 80%. The highest overlap between methods from different families was found between rankings based on dChip and RMA, with a median overlap of 70% over six datasets. The overlap between RMA and MAS is lower, with a median of only 56% over all six datasets. From the last block, we can see that even though dChip and mmgMOS are both multi-chip preprocessing strategies, they usually tend to pick different feature sets, with a median of 44% over six studies. Excluding the (mgMOS,mmgMOS) pair, the median overlap over all data sets and splits is 52%. Note that this lack in overlap is completely due to the preprocessing method chosen, as the feature selection criterion, the array platform, and the set of samples (and hence the sample handling and hybridization conditions) are all identical. Comparing the overlap from Figure 2.4A (preprocessing variability) to that in Figure 2.3 (perturbation variability) we see that the scores have a similar range, i.e. around 50%.

Relative strength of Affymetrix-based breast cancer signatures is more robust against preprocessing variability

In the previous section, we saw that the use of a different preprocessing strategy typically leads to the identification of a different feature set and that the overlap between top-ranked feature sets for different preprocessing pairs can be quite poor. Figure 2.4B shows the distribution of the relative strengths for top-ranked feature lists from the example in Figure 2.4A. The order of the boxplots in panel B is the same as in panel A. Comparing the two panels, we see that a lower overlap is typically associated with a lower relative strength as well. However, although the overlap between top-ranked features sets can be quite poor, the relative strengths are reasonably high. The highest scores are again obtained between preprocessing pairs from the same family. Since the (m)mgMOS models have a large overlap in top-ranked lists, their relative strengths are high as well, with values of over 90%. Even for the Loi dataset, the median relative strength over 50 splits is still above 89%, while the actual overlap is quite poor with a median of 60%. Furthermore, distributions of relative strengths for the Minn dataset, for pairs of preprocessing strategies from different families (all blocks except the first one), are mostly wider and have a lower tail than the other distributions. This is probably caused by the small number of samples in the Minn dataset. Comparing the relative strengths from Figure 2.4B (preprocessing variability) to those in Figure 2.3 (perturbation variability) we see that the scores are similar, with a mean relative strength of 85.1% taken over all entries in Figure 2.4B to a mean relative strength of 84.2% taken over all entries corresponding to Figure 2.3 at $n = 100$.

Preprocessing-neutral top gene lists

The lack of overlap between top-ranked lists corresponding to different preprocessing methods, as observed in Figure 2.4A, presents an additional complication in comparing performances between signatures based on such lists as we then cannot know whether a difference in performance is due to a difference in selected features, or due to a difference in feature values as obtained from the preprocessing method. In order to compare the performances of signatures constructed on data from different methods, ideally we would like to use the same set of features. Here, we show that we can obtain a ranking with a high relative strength over all preprocessing methods by combining the ranking information associated with the different preprocessing methods. In the previous sections, each top-ranked feature set was based on data from a single preprocessing method. For a given method $m \in M$ we will refer to this ranking as a *single-rank* feature list. The strength of a feature i for method

m, denoted by $S_m(i)$, was measured by $\text{SNR}_{i,m}$. Here we base the strength of feature i on the average of the individual strengths, as obtained by the different preprocessing methods in M, i.e., we use a strength

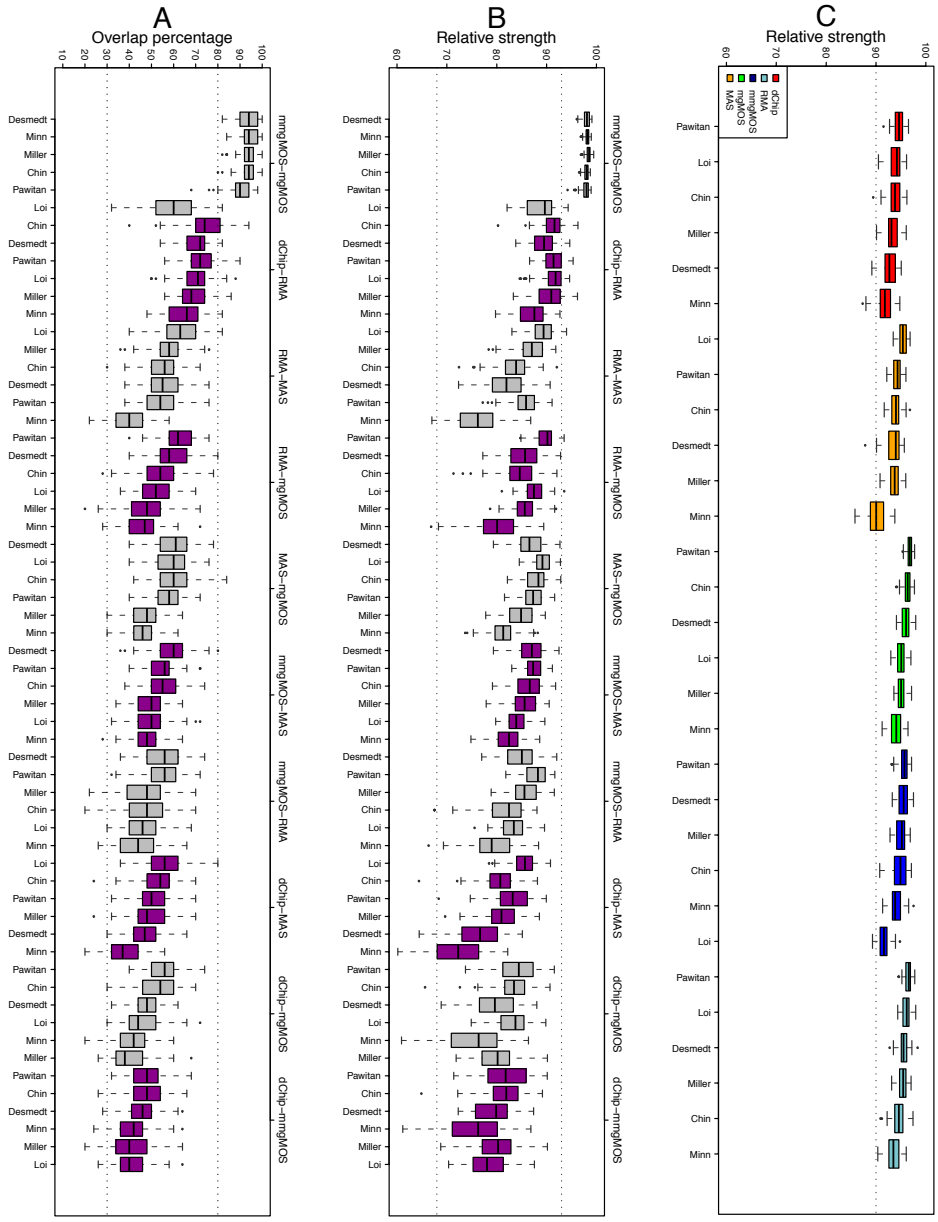
$$S(i) = \frac{1}{|M|} \sum_{m \in M} S_m(i). \tag{2.15}$$

In the remainder we will refer to the ranking based on this combined strength $S(i)$ as a *multi-rank* list. For each split k and for each dataset, we computed the top-100 ranked feature list based on this multi-rank strategy and determined its relative strength in the top-rank list $F_{100,m,k}$ for each preprocessing method m. Figure 2.4C gives for each dataset the distribution of these relative strengths. Relative strengths of the multi-ranked lists are high, with a median score of over 90% for all datasets. In order to decouple the effect of feature selection from the impact of perturbation and preprocessing variability on classification performance, we will therefore mainly use multi-rank gene lists, although all experiments on the Affymetrix datasets were also performed using the single-rank lists.

2.4.2 Impact of feature variability on classification

We start our investigation of the effects of feature variability on classification by taking an in-depth look into the van't Veer *et al.* (2002) and van de Vijver *et al.* (2002) expression data, which is based on the Rosetta error model. Starting from a single split of the data and using only features as considered in the original publications, we progress towards a more sophisticated setting, ending up in using the full sensitivity analysis protocol and applying it on all Affymetrix datasets using multiple preprocessing strategies and multiple classifiers.

Figure 2.4 ***(facing page)*****: Comparison of top-100 ranked features lists** $F_{100,m,k}$ and $F_{100,m',k}$, as obtained using different preprocessing strategies m and m', for different splits k. A. Percentage of the top-half of one list that is found in the other list, and vice vera. Each boxplot represents the distribution of such percentages over 50 splits, for a specific pair (m, m') (indicated on top of the figure). For each split, we determine the percentage of $F_{50,m,k}$ found in $F_{100,m',k}$ and the percentage of $F_{50,m',k}$ found in $F_{100,m,k}$. Each distribution thus contains $50 \cdot 2 = 100$ points. All boxplots corresponding to the same preprocessing pair are colored similarly. In total there are 15 distinct pairs. The pairs are ordered by the observed median overlap over all six datasets. B. Distributions of the relative strength scores for top-ranked feature lists corresponding to the various preprocessing pairs. C. Relative strength of the top-100 multi-ranked gene lists with respect to the original rankings, for each preprocessing method and each Affymetrix dataset.


A
Overlap percentage
B
Relative strength
C
Relative strength
dChip
RMA
rmngMOS
mgMOS
MAS
rmngMOS−mgMOS
dChip−RMA
RMA−MAS
RMA−mgMOS
MAS−mgMOS
rmngMOS−MAS
rmngMOS−RMA
dChip−MAS
dChip−mgMOS
dChip−rmngMOS
Desmedt
Minn
Miller
Chin
Pawitan
Loi

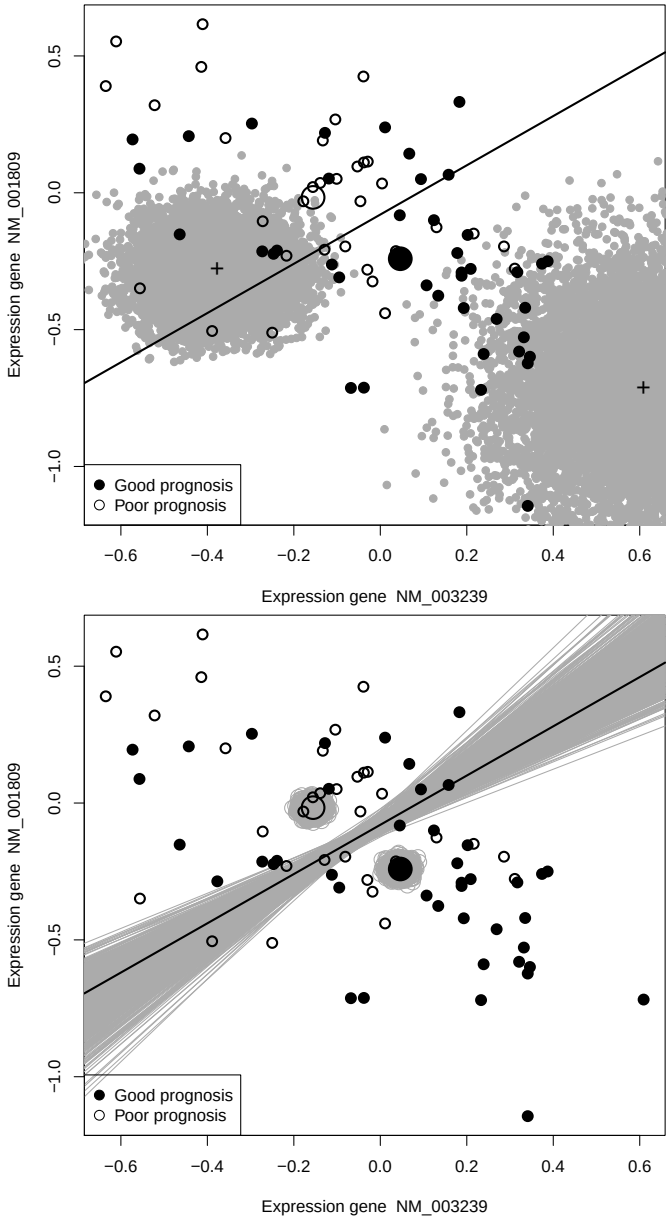


Figure 2.5: Examples on the influence of noise. The top panel shows the impact of noise on the classification of individual samples. Small circles indicate individual samples; the two large circles indicate class centroids. For two samples, indicated by small plus signs, we superimposed a 1000 perturbed instances, indicated by the gray clouds. The bottom panel shows the impact of noise on the estimation of the decision boundary, based on one 1000 perturbed instances of the training set. The gray clouds around the centroids indicate their variance, and the gray area around the decision boundary indicates its position for the one 1000 perturbations.

Rosseta two gene toy example

To illustrate the effect of measurement noise on classification, we consider a simplified setting by constructing a classifier with only two genes out of the 70-gene profile of van't Veer *et al.* (2002). Consider Figure 2.5. The top panel illustrates the impact of noise on two individual cases using a fixed decision boundary. Samples are classified corresponding to the closest of the two class centroids (average profiles), resulting in the decision boundary as depicted in the figure.

In the top panel we see that the left-hand cloud has points on either side of the decision boundary, indicating that the classification of the corresponding sample is unstable. The right-hand cloud is completely on one side of the boundary, indicating a stable assignment. The lower panel of Figure 2.5 shows the impact of noise on the estimation of the class centroids and thus on the decision boundary itself. This is derived by perturbing all training samples 1000 times and re-estimating the decision boundary. Although the centroids are less noisy than individual samples, this still leads to unstable classifications of a few samples.

Van 't Veer signature is sensitive to perturbation variability

We investigated the classification stability of the original 70-gene signature of van't Veer *et al.* (2002). The classifier used for the construction of their signature is a nearest centroid classifier. Classification for this classifier can be linked to a *discriminant score* (see Materials and Methods), by which we assign a sample to the good prognosis class if the discriminant score is positive, and to the poor prognosis class otherwise. We use the original Van 't Veer training set of 78 samples to estimate the class centroids. As a validation set we took the 106 remaining samples in the Rosetta dataset. Next, using the uncertainty information estimated by the Rosetta error model (see Materials and Methods), we created one 1000 perturbed versions of the validation set and classified these with the classifier built on the original training data.
Figure 2.6 shows the impact of perturbation variability on the discriminant scores for each of the 106 cases. Note that a validation sample is stably classified if the discriminant score is either positive for all its perturbed instances, or negative for all its perturbed instances. For some samples the variation of the corresponding discriminant score is small, while for others it is quite large, reflecting the fact that measurements for the same probe on different arrays are associated with different measurement errors. In addition, the individual distributions are quite symmetric, which stems from the fact that the classifier is linear and we added symmetrical noise. Perturbation variability can indeed

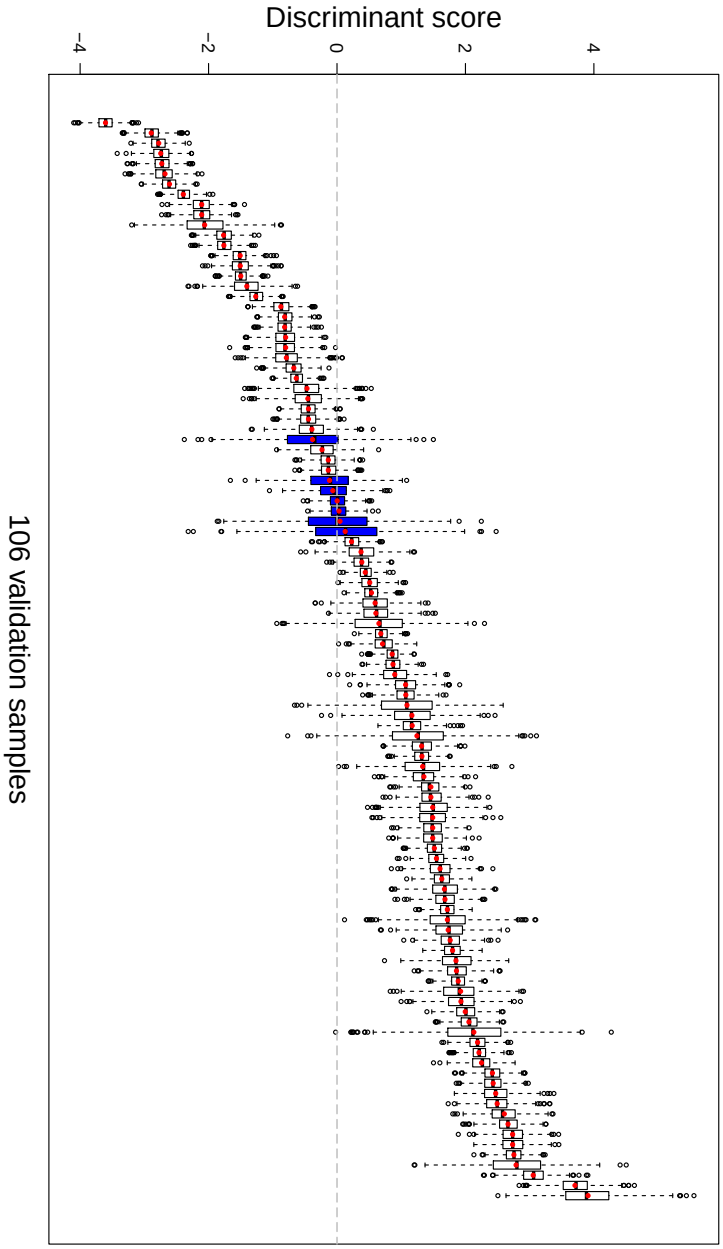


Figure 2.6: Discriminant score $\mathbf{x}^T\mathbf{w}$ for each of the 106 validation samples, when using a nearest centroid classifier built on the 70-gene profile of van't Veer *et al.* (2002), over 1000 perturbations. Perturbed expression data is based on the Rosetta error model. Red dots indicate the discriminant scores corresponding to the unperturbed expression data. The blue boxes indicate samples with a map-score of at least 25%.

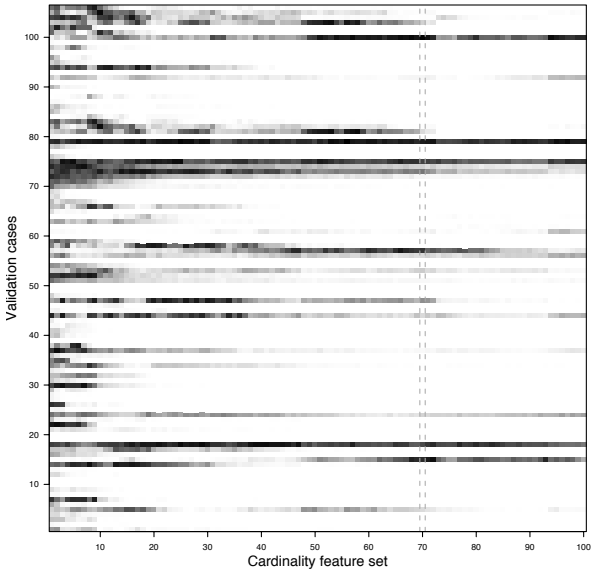


Figure 2.7: Map matrix example. The minimum assignment percentages (white = 0%, black = 50%) for the 106 validation samples and signatures of increasing size, determined over 1000 perturbations of the validation data. The column indicated by the dashed lines corresponds to the original 70-gene signature. The figure shows the impact of noise on the validation cases using a fixed decision boundary c.f. top panel Figure 2.5.

disrupt the classification process, since for seven samples (indicated in blue) the box-and-whisker plots cross the horizontal threshold line at height zero. Note that the boxes in a box-and-whisker plot indicate the interquartile range of a distribution and thus these seven samples have an associated map-score of at least 25%.

A map-matrix example for the Rosetta dataset

Now we extend the example of the previous section by considering a sequence of 100 signatures constructed using the top-100 ranked features from the Van 't Veer data and zoom in on the impact of perturbation variability on classifications of individual samples from the Rosetta data by taking an in-depth look at a map-matrix, such as the ones obtained from our sensitivity analysis protocol. Classifications are again performed using the nearest centroid classifier. The n^{th} signature is constructed using only the top-n features. Note that this setting is similar to our protocol, in which at step 2 we take the 78 training cases of Van 't Veer data as a training set, the 106 remaining samples as a validation set, at step 3 take the top-100 features as described above and at step 4 train a sequence of 100 NC classifiers, thus yielding 100 signatures.

Following the protocol, at step 7 we obtain a map-matrix, which in this case is a 106 by 100 matrix, where the entry at row k and column n contains the map-score of sample k using a signature involving the top-n features.
Figure 2.5 visualizes the map-matrix of this example by means of a heatmap. Here white entries indicate completely stable assignments, i.e. the map-score is zero, while black entries indicate random class assignments. From the figure we see more dark areas on the left than on the right, indicating that classification is generally less stable if fewer genes are used. In addition, for very small signature sizes i.e. less than 10, the classification of virtually all samples can be disrupted by perturbation variability, as almost none of the corresponding cells are completely white. Furthermore, we observe that for some samples, adding features may first reduce the impact of variability, whereas adding more features later increases the impact of variability again and vice versa. This may be due to the fact that either features are added that are quite noisy for such a sample, or that such features draw these samples closer to the decision boundary. Finally, even for large signatures the classification of some samples can still be affected by perturbation variability, although the number of such cases is typically low.

Performance and stability curves for the Rosetta dataset

In the previous section, results were obtained using only a single split of the data in a training and validation set. Here we apply the full sensitivity analysis protocol to the Rosetta dataset consisting of 184 samples. Figure 2.8 shows the resulting performance (P) and stability (S) curves for five classifiers, based on 50 splits of the data. The NC classifier performs best and clearly increases its performance when using more features with a highest performance of around 65%. This is comparable to the estimates reported in Wessels *et al.* (2005); Michiels *et al.* (2005); van Vliet *et al.* (2008). Furthermore, on this dataset the NC classifier also had the best S-curve. S-curves generally improve when using more features, however, none are flat and close to zero, indicating that perturbation variability can consistently disrupt these classifications. For the NC classifier the impact of perturbation variability on this dataset quickly diminishes, with an average number of unstable assignments leveling off around only 2.5% at a signature size of 100. For other classifiers we see that the impact of perturbation variability is higher than for the NC classifier and especially the 1-nearest neighbor seemed very sensitive at small signature sizes, only leveling off around 10% at a size of 100 features. Although stability is a desirable characteristic, we should not directly link ascending P-curves to descending S-curves and simply attribute the higher performance of the NC classifier to perceived noise tolerance. Although the S-curves typically decrease when the signature size increases, the P-curve does not generally

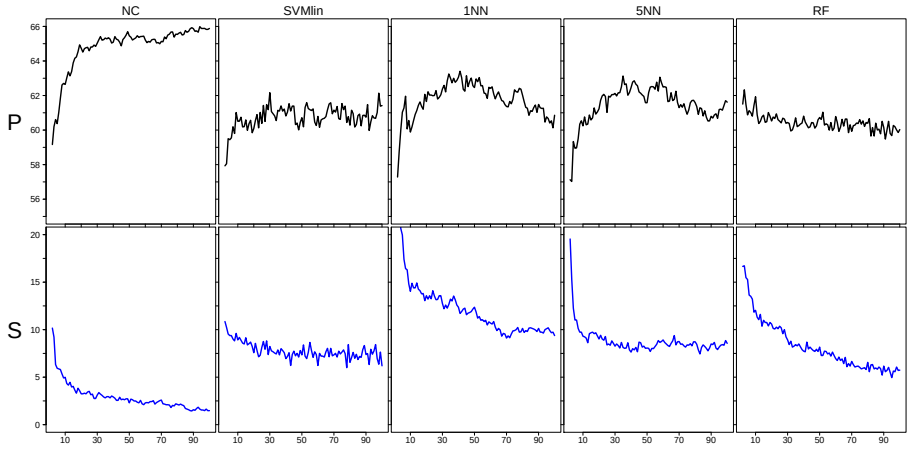


Figure 2.8: Performance and stability curves for the Rosetta dataset, for different classifiers. Results are based on averages over 50 splits, S-curves were created using a threshold of 35. The x-axis shows the signature size, while the y-axis shows the average balanced accuracy rate over 50 splits.

show such a monotonic behavior. For instance, the nearest neighbor classifier shows a decreasing P-curve for larger signature sizes and is indeed known to be intolerant to the inclusion of irrelevant features.

Impact of feature variability for Affymetrix datasets

In order to investigate the impact of both preprocessing variability and perturbation variability on Affymetrix GeneChip data, we ran our complete protocol, for each of the six Affymetrix datasets, five different preprocessing methods (mgMOS, mmgMOS, MAS 5.0, dChip, and RMA), and six classifiers (NC, 1NN, 3NN, SVMlin, SVMrbf, and RF). Each dataset was analyzed with 50 different splits into a training and validation set. Each validation set was perturbed 1000 times in order to infer the S-curves. Furthermore, the experiments were performed using both the single-rank and multi-rank sets. From Figure 2.4A we saw that different preprocessing methods tend to pick different features and that the overlap between rankings can be low. Hence, if we use single-rank sets, we will observe a combined effect where differences between curves corresponding to different preprocessing methods can be due to a difference in signature composition, as well as due to feature variability. The advantage of the multi-rank approach is that for a given dataset-classifier pair, observed differences in performance, discordance (D, see Methods section), and stability curves are not due to a difference in signature composition, but solely due to feature variability. Using a signature based on a multi-rank set

effectively decouples the impact of feature selection from the effect of feature variability on classification performance. Given the high relative strengths of the multi-rank sets, as observed in Figure 2.4C, we therefore show in the main text only the figures corresponding to these multi-rank sets. Figures 8, 9 and 10 show the resulting P, D and S-curves, respectively, for all 36 classifier-dataset combinations. The corresponding P, D and S-curves for the single-rank sets are shown in Additional Files 4, 5 and 6 online, respectively.

Lack of overlap in performance curves on Affymetrix datasets

Figure 2.9 shows the P-curves for the multi-rank based experiments. Note that we are less concerned with the actual shape of the performance curves, but we are mainly interested if different preprocessing methods lead to overlapping curves. For most dataset-classifier pairs the corresponding P-curves indeed show the same trends. In direct contradiction to our null hypothesis, however, several large deviations can be seen, most notably on the Pawitan dataset for multiple classifiers (NC, 1NN, 3NN, RF). On this dataset there seems a clear advantage in using dChip or RMA expression estimates. Although RMA usually performs well, it does not consistently give the best performance curves. In fact, no preprocessing method is clearly superior to all other methods. On the datasets of Desmedt and Minn, for instance, MAS often outperforms both RMA and dChip. On most datasets balanced accuracy rates between 60 and 75% could be achieved, depending on the classifiers and signature size. In most cases the performance increases for larger sized signatures, although on the dataset of Desmedt high accuracies could be achieved using only a few features. Although certain preprocessing-classifier pairs have a good performance for some datasets, such performance advantages cannot be maintained on the other datasets. When comparing classifiers, we see that simple classification models like the NC classifier and the NN classifiers typically perform at least as well as more complex classifiers like SVM or RF. Similar observations on the performance of simple versus more complex classifiers in the context of microarray data have been made in Dudoit *et al.* (2002a); Wessels *et al.* (2005); van Vliet *et al.* (2008). In our experiments the SVM classifiers did not perform well. For instance, even though SVMlin and NC are both linear classifiers, the NC classifier is clearly superior. Although a large grid of hyperparameter values was attempted for SVM, it proved hard to find the correct hyperparameter values.

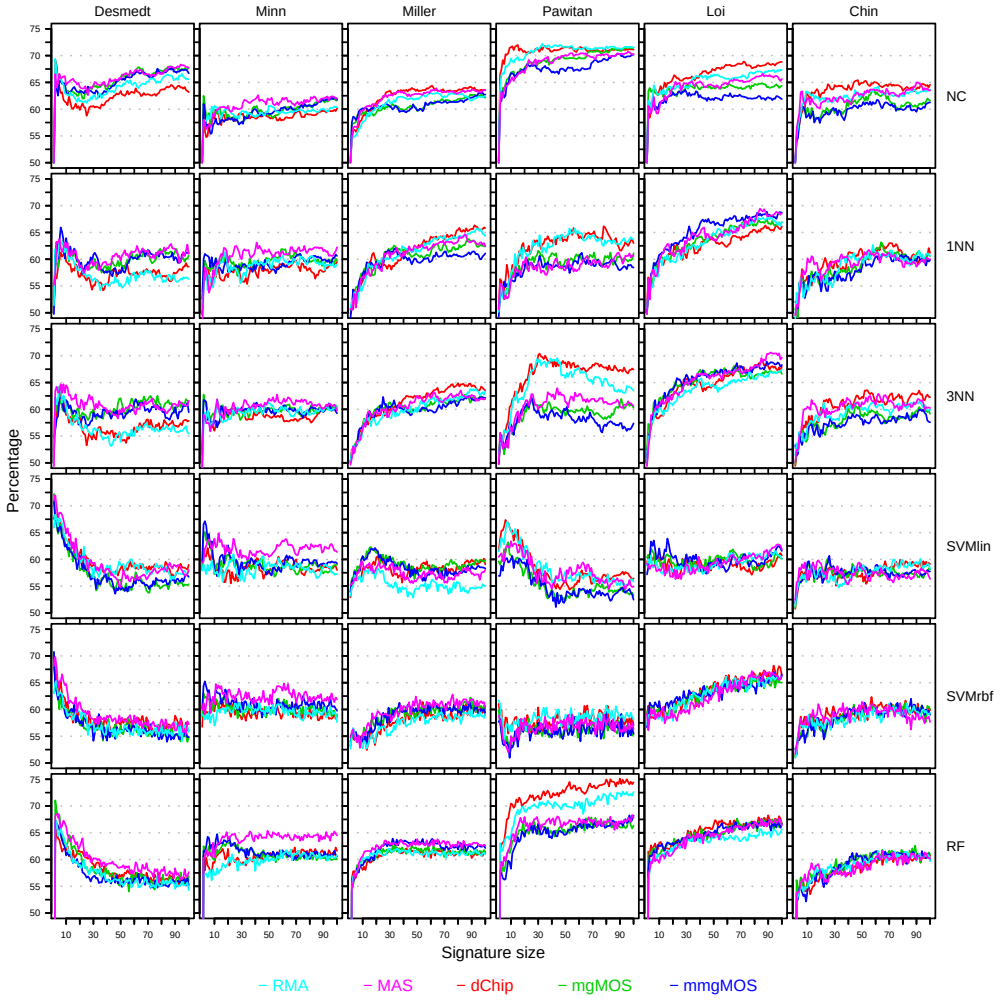


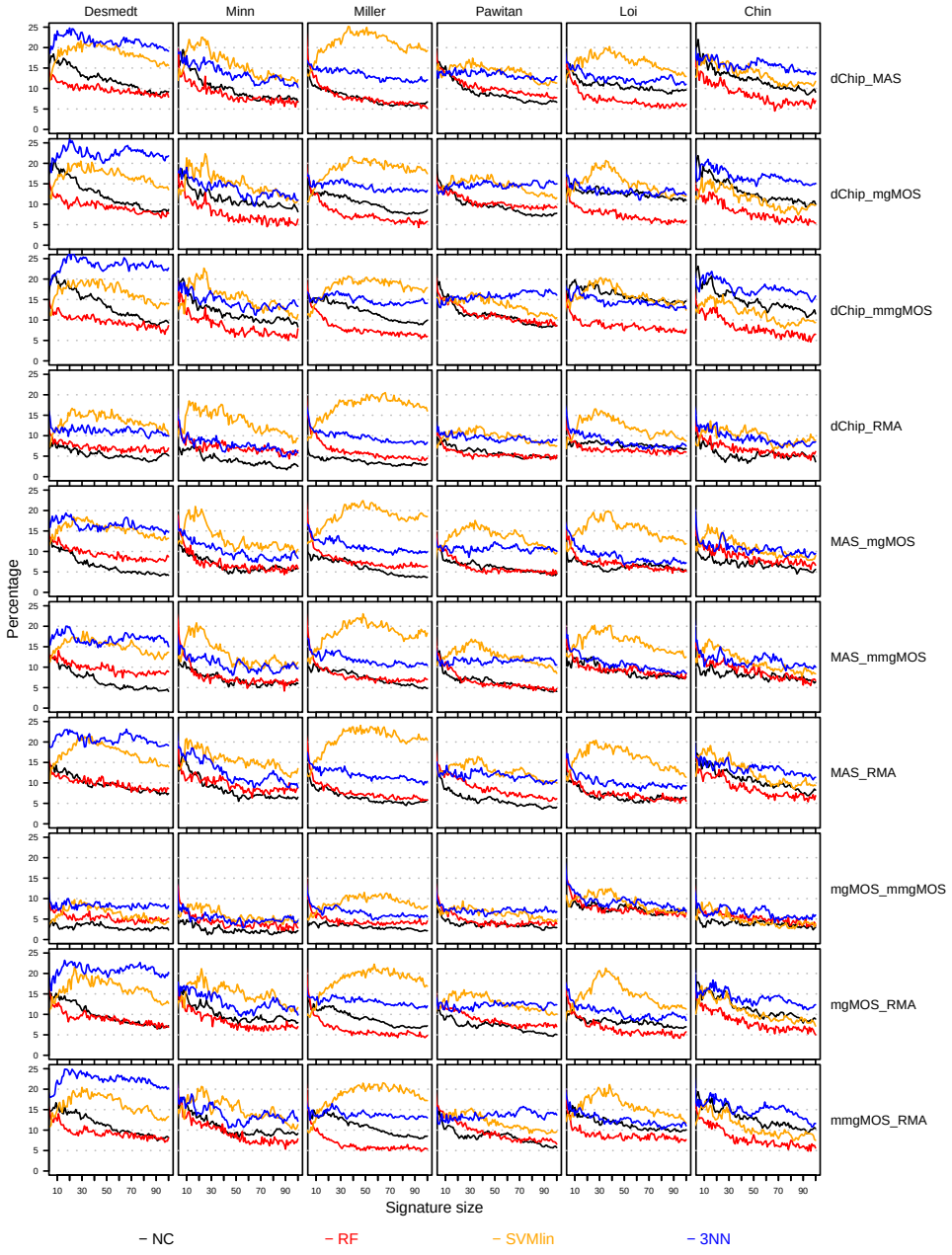
Figure 2.9: Performance curves for the Affymetrix datasets, for different preprocessing methods, and classifiers. Rows represent curves obtained using different classifiers, while columns represent curves for different datasets. Within each cell, performance curves associated with different preprocessing methods are shown in separate colors. The color scheme is shown at the bottom of the figure. Within a cell the x-axis provides the signature size, while the y-axis gives the average balanced accuracy over 50 splits. For each dataset and split, the top-100 feature set was computed using the multi-rank strategy of Section 2.4.1 and this ranking was subsequently used for all classifiers in order to construct signatures.

Different preprocessing methods produce discordant class assignments

In the previous section we observed that in several studies there was a lack in overlap between performance curves for different preprocessing methods, clearly indicating a discordance in outcome prediction. Even in the case of overlapping performance curves, however, one cannot ascertain that the individual class assignments are concordant. Figure 2.10 shows for several classifiers (NC, 3NN, SVMlin, RF) the discordance curves corresponding to the P-curves of Figure 2.9. For all preprocessing pairs clear discrepancies can be seen, which is in direct disagreement with our null hypothesis. The highest D-curves for the selected classifiers are observed for the 3NN and SVMlin classifier, with an overall median discordance (over all signature sizes and splits) of 12.6% and 14.2%, respectively. The NC and RF classifiers show lower numbers of discordant class assignments with an overall median discordance of 8.4% and 7.5%, respectively. For the latter two classifiers the discordance also clearly decreases with larger signature sizes, leveling off at a signature size of 100 with an overall median discordance of 6.8% and 6.4%, respectively.

The discordance is often larger in the poor prognosis group than in the good prognosis group (see Additional Files 7 and 8 online, respectively. Note that in most breast cancer datasets, the former group is also much smaller than the latter. When using balanced performance indicators, a discordance in the poor prognosis group is then more heavily penalized than a discordance in the good prognosis group. For instance, in Figure 2.9 a clear difference in performance curves can be seen for the preprocessing pair (dChip,MAS), when applying the RF classifier on the Pawitan dataset. From Figure 2.10, however, the lack in concordance for the preprocessing pair (dChip,MAS) does not seem much larger than on other datasets. From Additional File 7 online, we can see that for this preprocessing pair and dataset the number of discordant cases for the RF classifier in the poor prognosis group is indeed higher, with an overall median of 18.7% compared to an overall median of 8.5% on the remaining datasets.

Figure 2.10 ***(facing page)*****: Discordance curves for the Affymetrix datasets, for all 10 distinct preprocessing pairs, and four classifiers.** Rows represent different preprocessing pairs, while columns represent curves for different datasets. Within each cell, discordance curves corresponding to different classifiers are shown in separate colors. The color scheme is shown at the bottom of the figure. Within a cell the x-axis provides the signature size, while the y-axis gives the average percentage of cases, over 50 splits, of inconsistent class assignments on the unperturbed validation sets. For each dataset and split, the top-100 feature set was computed using the multi-rank strategy of Section 2.4.1 and this ranking was subsequently used for all classifiers in order to construct signatures.


Desmedt
Minn
Miller
Pawitan
Loi
Chin
dChip_MAS
dChip_mgMOS
dChip_mmgMOS
dChip_RMA
MAS_mgMOS
MAS_mmgMOS
MAS_RMA
mgMOS_mmgMOS
mgMOS_RMA
mmgMOS_RMA
Percentage
Signature size
NC
RF
SVMlin
3NN

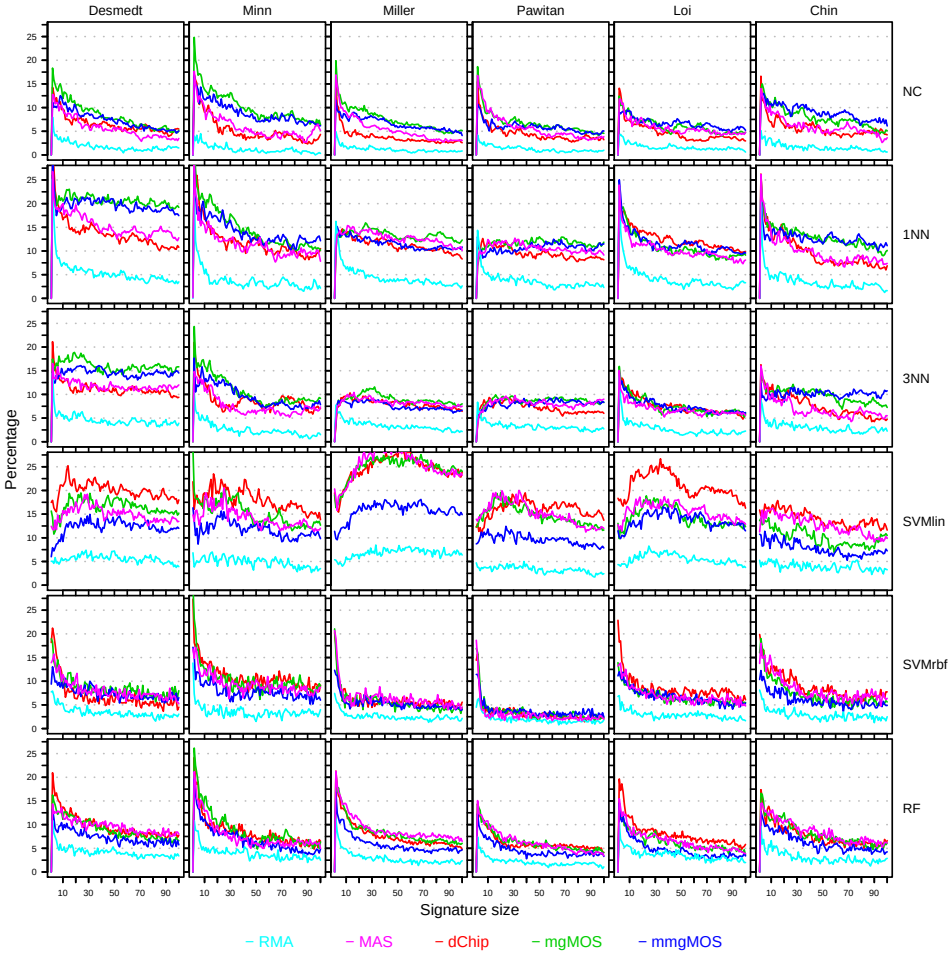


Figure 2.11: Stability curves for the Affymetrix datasets, for different preprocessing methods, and classifiers. Rows represent curves obtained using different classifiers, while columns represent curves for different datasets. Within each cell, stability curves associated with different preprocessing methods are shown in separate colors. The color scheme is shown at the bottom of the figure. Within a cell the x-axis provides the signature size, while the y-axis gives the average percentage of cases over 50 splits with a MAP-score larger than 35. For each dataset and split, the top-100 feature set was computed using the multi-rank strategy of Section 2.4.1 and this ranking was subsequently used for all classifiers in order to construct signatures.

High impact of perturbation variability for small signature sizes on Affymetrix datasets

Figure 2.11 shows the stability curves associated with the class assignments of Figure 2.9. None of the S-curves are flat and located near zero, which is again in direct contradiction with our null hypothesis. For most classifiers and preprocessing methods the impact of perturbation variability is high at small signature sizes, in which over 10% of the assignments are unstable. Similarly to Figure 2.8, the impact of perturbation variability quickly diminishes for increasing signature sizes, although for most classifiers approximately 5% of the assignments are still unstable at a signature size of 100. The perturbations corresponding to RMA appear to be smaller compared to those of the other preprocessing methods, as RMA consistently gives the lowest S-curves. These S-curves cannot always be associated with the best P-curves though. When comparing classifiers we see that the impact of perturbation variability can be quite different for different classifiers. Certain classifiers like SVMs (Statnikov *et al.*, 2008) and RF (Breiman, 2001) have been claimed to be noise tolerant. We did not find clear evidence that SVM or RF are more tolerant to the types of perturbation variability as discussed here. Although the SVMrbf indeed appears very stable on some datasets, its performance is also very poor compared to other models (Figure 2.9). The S-curves corresponding to SVMlin are notably different and the class assignments seem particulary sensitive to perturbation variability. No satisfactory answer was found that could explain this observed behavior. Furthermore, in our experiment the RF classifier is not more noise tolerant than for instance the NC classifier. For small signature sizes, i.e. fewer than 10 genes, the average number of unstable assignments (taken over all studies and all preprocessing methods except RMA) is 11.8% for RF, compared to only 10.1% for the NC classifier. At a size of 100, the average number of unstable assignments for RF and NC is 5.3% and 4.6%, respectively. Finally, the impact of perturbation variability for the nearest neighbor classifiers appears to be larger. For 1NN and 3NN the average number of unstable assignments at signature sizes less than 10 is 15.5% and 11.3%, respectively, and at size 100 it is 10.6% and 8.8%, respectively.

2.5 Discussion

Finding high-quality stable biomarkers in breast cancer applications using microarray expression profiling has proven to be quite challenging with reported balanced accuracy rates for most breast cancer signatures somewhere between 60 and 70%. Signature composition strongly depends on the subset of patient samples used for feature selection (Ein-Dor *et al.*, 2005). Furthermore, a high

level of inconsistency between individual class assignments between different signatures has been reported (Reyal *et al.*, 2008). Differences in array platforms as well as biological variability have been conjectured to play a major role in these discrepancies.

We designed an experimental protocol to evaluate the impact of two other types of variability, namely preprocessing and perturbation variability, on signature composition and classification. For this purpose several state of the art and frequently applied preprocessing methods were selected. Complementary to Ein-Dor *et al.* (2005), we showed that signature composition is strongly influenced by perturbation variability and preprocessing variability, even if the array platform and the stratification of patient samples are identical. In addition, using our multi-rank feature sets we showed that there is often a high level of discordance between individual class assignments for signatures constructed on data coming from different preprocessing schemes, even if the actual signature composition is identical. For the single-rank feature sets, the observed discrepancies were even larger. No preprocessing scheme, however, yielded data that was clearly superior for classification purposes. When comparing preprocessing variability to perturbation variability, we found their impact on feature selection to be equally strong. On classification, however, the impact of preprocessing variability often remained strong with increasing signature size, whereas the impact of perturbation variability quickly diminished.

Preprocessing noise is mainly caused by different underlying assumptions that are made on the data and on the available sources of information that are used. Some methods deliberately ignore some sources of information or exclude certain steps. RMA, for instance, does not use mismatch probe information to infer expression levels, while standard applications of dChip do not perform a background correction step. Note that the latter can have great implications on the final expression data for both one (Bolstad, 2004) and two-color array data (Ritchie *et al.*, 2007).

Our stability analysis performs explicit noise tolerance tests for a diverse set of classification routines, by using the class assignments of perturbed expression profiles. The results indicate that all classifiers considered were sensitive to perturbation variability, although the impact was much stronger at small signature sizes and quickly diminished for larger signature sizes. Furthermore, in most cases we found the level of noise tolerance for the NC, SVMrbf, and RF classifiers to be very comparable.

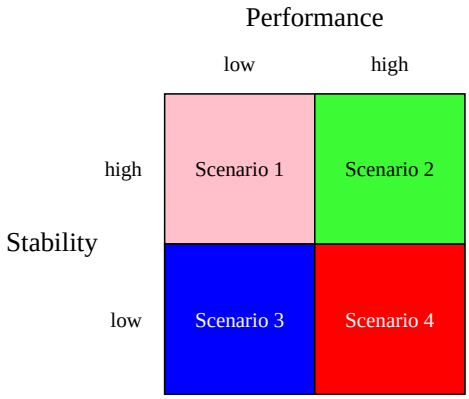


Figure 2.12: Different scenarios in the trade-off dilemma of performance vs. stability.

We chose to use realistic estimates of gene-wise measurement error in the stability analyses. Methods like the Rosetta error model, but also the mgMOS and mmgMOS models, are specifically designed to obtain such uncertainty information associated with the fitted expression data. Methods like dChip, RMA, and MAS 5.0 are not designed with this goal in mind. However, some uncertainty information can be derived from the summarization step, as performed in the preprocessing cascade. Although the uncertainty estimates for dChip, RMA and MAS are based on the same type of information, we found that perturbations corresponding to RMA seemed much less severe than those based on other methods; cf. Irizarry *et al.* (2006). For the Affymetrix preprocessing methods a potential problem with basing uncertainty estimates solely on the summarization step is that most probesets consist of a small number of probes, with a median size of 11 for the GeneChips used here, which can make the standard error estimates less reliable. Although the stability curves for MAS and dChip were closer to those of the mgMOS and mmgMOS models, from our experiments one cannot tell if RMA underestimates the errors or if the other methods overestimate the errors.

Our results also show that a high stability and a good performance do not always go hand in hand (Figures 2.8 and 2.10). Although stability is a desirable property, it is sometimes conflicting with achieving a high performance, which presents us with a dilemma, similar to the bias-variance dilemma (Duda *et al.*, 2001). To this end, consider Figure 2.12. From a classification standpoint, the second scenario is obviously the preferred scenario, while scenario three is equal to tossing a coin. Note that scenario one can always be achieved by using a rule that assigns all samples to the same class. Such a rule is

extremely stable, yet when using balanced accuracy rates, will also have a poor performance. This scenario was sometimes observed for the SVM classifiers. For both linear and non-linear SVMs, parameter estimation was hard. This might be an explanation for the observed poor performance, although our performance estimates on single datasets for SVM were often comparable to those reported earlier (van Vliet *et al.*, 2008; Kim, 2009). Finally, scenario four would be a strong indicator that the perturbed expression profiles are not very realistic, given the fact that performance and stability are both measured on the same validation data. This scenario was, however, not observed in our experiments. We did encounter this scenario in attempts to base perturbed expression profiles on jitter i.e. artificial noise estimates. The main problem in using jitter is that such estimates are either much too low or much too high and therefore this type of perturbation was not further pursued here.

Note that our goal was not to compare classifiers or even to find optimal biomarkers per se and it is likely that the performance of some classifiers can be further improved e.g. by changing the feature selection step in our protocol, which in our case was based on univariate signal-to-noise-ratio statistics. Alternative univariate ranking strategies such as those based on the t-test, Mann-Whitney u-test, and Mahalanobis distance were reported to perform similarly (Wessels *et al.*, 2005) and were therefore not pursued here. Note that the former methods all construct rankings based on binary class-label information. Survival information on which the class labels are based could be incorporated in the ranking step as well. For instance, in Wang *et al.* (2005) a 76-breast cancer gene signature was derived using a ranking step based on information from univariate Cox proportional-hazards regression models using the length of distant metastasis free survival.

For some classifiers it might be advantageous to resort to multivariate wrapper-based feature selection methods. Perhaps the simplest computationally efficient multivariate wrapper is the Top-Scoring-Pair (TSP) classifier (Geman *et al.*, 2004), which performs its classifications on the basis of the expression values of just two genes. On several classical tumour data sets e.g. leukemia (Golub *et al.*, 1999), colon (Alon *et al.*, 1999), lymphoma (Shipp *et al.*, 2002), and prostate (Singh *et al.*, 2002), the TSP was able to obtain balanced accuracy rates well over 90%; see Tan *et al.* (2005). Note that the TSP is a rank-based method, that is characterized by replacing expression levels by their corresponding ranks. In Geman *et al.* (2004), the TSP is claimed to be invariant to pre-processing changes, as it is invariant to any monotonic transformation of the expression data. Although our forms of feature variability are very realistic, they can certainly not be considered as monotonic transformations of the raw

expression data. Initial experiments with the TSP on the Affymetrix breast cancer datasets revealed that the classifier was extremely sensitive to feature variability, with corresponding balanced accuracy rates often close to 50% (data not shown). On the Rosetta data a similar observation on the performance of the TSP was reported in Lai *et al.* (2006). As the TSP uses only two genes, these results are in agreement with our observation that breast cancer signatures comprised of few genes seem very susceptible to feature variability. In addition, in Lai *et al.* (2006) several alternative multivariate approaches were benchmarked, with the main conclusion that multivariate selection approaches often do not lead to consistently better results than univariate approaches. Moreover, compared to multivariate approaches, univariate ranking procedures have the benefit of a considerable computational speed up, which in our case was very important considering the large number of experiments performed.

Our sensitivity analysis was performed on a sizable collection of patient sample hybridizations and in a breast cancer classification context, which is different from the small scale spike-in and dilution studies on which most previous microarray sensitivity analyses were performed (McCall and Irizarry, 2008; Cope *et al.*, 2004). One advantage of the latter two types of studies is that the ground truth is known, which for most breast cancer studies is less obvious. In our framework, however, under the null hypothesis we also know exactly what should be expected, i.e. for different preprocessing methods or for perturbed versions of a dataset we should have selected the same features, had overlapping P-curves and obtained D-curves and S-curves that were zero for all signature sizes, as stated in our null hypothesis. Based on the outcome of our experiments, however, we conclude that this is not the case, and hence we conclude that in microarray *breast cancer* studies feature variability can have a strong impact on both feature selection and classification. We conjecture feature variability to be less of an issue in microarray studies for which a high performance can be obtained such as for the classical tumour datasets mentioned above. Note that these studies all deal with tissue-type recognition problems, which are considerably easier classification problems than event prediction studies, such as the breast cancer studies treated here; see also Wessels *et al.* (2005).

Finally, the work presented in this chapter was mainly of a descriptive nature, analyzing the impact of feature variability. Obviously, one would next like to enhance the performance and stability of classifiers by exploiting the feature variability information. For instance, in the context of point injection techniques, one could use the perturbed expression profiles as additional candidates to be injected, instead of the rather artificial candidates obtained

by linear interpolation (Dudoit *et al.*, 2002a). Another avenue that one may take is to directly increase classification concordance by explicitly enforcing it, for instance in a wrapper framework.

2.6 Conclusion

We performed an extensive sensitivity analysis of microarray breast cancer classification under feature variability. Our results indicate that signature composition is strongly influenced by preprocessing variability and perturbation variability, even if the array platform and the stratification of patient samples are identical. In addition, we show that there is often a high level of discordance between individual class assignments for signatures constructed on data coming from different preprocessing schemes, even if the actual signature composition is identical.

We presented evidence of discrepancies induced by technical variation that cannot be considered negligible, as previously claimed by some researchers (Klebanov and Yakovlev, 2007). We therefore strongly recommend that feature variability is taken into account during the construction of a signature, especially when using microarray technology for the classification of individual patients. In addition, measures should be taken to minimize the technical variation of microarray procedures when used for such high impact applications as cancer diagnostics.

CHAPTER 3

BREAST CANCER SUBTYPE PREDICTORS REVISITED: FROM CONSENSUS TO CONCORDANCE?

3.1 Abstract

Background At the molecular level breast cancer comprises a heterogeneous set of subtypes associated with clear differences in gene expression and clinical outcomes. Single sample predictors (SSPs) identify subtypes via a two-stage approach consisting of clustering and predictor construction based on the cluster labels of individual cases. SSPs have been criticized because their subtype assignments for the same samples were only moderately concordant ($\kappa < 0.6$).

Methods We propose a semi-supervised approach where for five datasets, consensus sets were constructed consisting of those samples that were concordantly subtyped by a number of different predictors. Next, nine subtype predictors - three SSPs, three subtype classification models (SCMs) and three novel rule-based predictors based on the St. Gallen surrogate intrinsic subtype definitions (STGs) - were constructed on the five consensus sets and their associated consensus subtype labels. The predictors were validated on a compendium of over 4,000 uniformly preprocessed Affymetrix microarrays. Concordance between subtype predictors was assessed using Cohen's kappa statistic.

Results In this standardized setup, subtype predictors of the same type (either SCM, SSP, or STG), but with a different gene list and/or consensus training set, were associated with almost perfect levels of agreement (median

$\kappa > 0.8$). Interestingly, for a given predictor type a change in consensus set led to higher concordance than a change to another gene list. The more challenging scenario in which the predictor type, gene list and training set were all different resulted in predictors with only substantial levels of concordance (median $\kappa = 0.74$) on independent validation data.

Conclusions Our results demonstrate that, for a given subtype predictor type stringent standardization of the preprocessing stage, combined with carefully devised consensus training sets, leads to predictors that show almost perfect levels of concordance. However, predictors of a different type are only substantially concordant, despite reaching almost perfect levels of concordance on training data[1].

3.2 Introduction

In the last decade substantial advancements have been made in our ability to probe the human transcriptome, especially by high-throughput techniques such as microarrays and more recently by next generation sequencing, i.e. RNA-seq. These techniques have deepened our understanding of complex diseases such as breast cancer (Weigelt *et al.*, 2010d). Genome-wide studies have also firmly established the notion that breast cancer does not constitute a single disease at the molecular level, but comprises a heterogeneous set of subtypes, associated with striking differences in gene expression patterns, clinical outcome and response to therapy (Sotiriou and Piccart, 2007). One of the most widely adopted subtyping schemes in this regard is the one introduced by Perou *et al.* (2000), which distinguishes the subtypes luminal (subsequently divided in the subgroups A, B, and/or C), basal, HER2 and normal-like.
Subtypes have mainly been identified via a two-stage approach (Wang *et al.*, 2013). In the first stage an initial grouping of samples of the same subtype is identified by hierarchical clustering, i.e. by unsupervised learning. Important ingredients of such schemes are the linkage criterion, distance measure and feature list. In the context of subtyping, the latter is often referred to as the intrinsic gene list (IGL) (Perou *et al.*, 2000). In the second stage a predictor is constructed based on supervised learning: cluster labels of individual cases from the first stage are used as class labels in order to train a predictor, often of the nearest centroid type. In breast cancer literature these predictors are frequently referred to as single sample predictors (SSPs) (Weigelt *et al.*, 2010a). Note that once an SSP has been fitted, new cases can be subtyped without a clustering stage. The most well-known breast cancer SSPs are those by Sørlie

[1]Status: HMJ Sontrop, MJT Reinders, PD Moerland. Breast cancer subtype predictors revisited: from consensus to concordance? (2014). Submitted.

et al. (2003), Hu *et al.* (2006) and PAM50, developed by Parker *et al.* (2009). In the remainder we will refer to these three predictors as the classic SSPs.

The two-stage approach towards subtype identification is, however, not without its pitfalls. Weigelt and colleagues (Weigelt *et al.*, 2010a) reported a low concordance between subtype assignments by the classic SSPs on four single-channel and dual-channel microarray datasets. They conclude that the classic SSPs do not reliably assign subtypes to individual patients and that therefore such identifications are not ready yet for routine clinical practice. The study was strongly criticized by Perou *et al.* (2010) and Sørlie *et al.* (2010) based on bioinformatics-based technical limitations, claiming that the findings were flawed due to the use of uncentered data. In a subsequent rebuttal Weigelt *et al.* (2010c), however, showed that properly centering the data did not lead to substantial improvement of the levels of concordance. The findings by Weigelt *et al.* (2010a,c) were corroborated by a meta-analysis of a substantially larger number of datasets from a variety of microarray platforms (Haibe-Kains *et al.*, 2012). Herein, Haibe-Kains and colleagues reported low robustness and concordance for SSPs and proposed SCMGENE (Haibe-Kains *et al.*, 2012), a robust three-gene model based on the subtype classification model (SCM) methodology using a Gaussian mixture model on a set of module scores (Desmedt *et al.*, 2008).

From the findings of Weigelt *et al.* (2010a,c) and Haibe-Kains *et al.* (2012), an unsettling notion on the reliability of SSPs emerges. However, these studies have several limitations which may have negatively influenced the observed concordance. Firstly, concordance assessments were made on data from multiple platforms, often different from the one(s) on which the SSPs had originally been constructed. Secondly, they used publicly available expression data that had been normalized by a variety of normalization schemes, even for data from the same platform. Thirdly, the classic SSPs were not specifically designed to be concordant at the individual sample level (Perou *et al.*, 2010). Perou *et al.* (2010) present PAM50 as a logical evolution over time in which several deliberate design changes were made compared to previous versions such as the SSPs of Sørlie and Hu. In that perspective, one could even argue that the discordance of the classic SSPs does not actually present a problem.

Here, we attempt to unify the different and sometimes conflicting views expressed in the articles by Weigelt *et al.* (2010a,c), Perou *et al.* (2010), Sørlie *et al.* (2010) and Haibe-Kains *et al.* (2012). We do so by analyzing subtype predictors in a setup in which all predictors are specifically designed to be highly concordant at the individual sample level. For five training sets, a semi-supervised approach was used to construct corresponding consensus sets (CSs) consisting of those samples that were concordantly subtyped by a number of different predictors selected from three classes of subtype predictors: (i) the

PAM50 SSP, (*ii*) three re-fitted SCMs and (*iii*) a novel rule-based predictor (STG) based on the surrogate intrinsic subtype definitions proposed at the 2011 St Gallen Consensus Conference (Goldhirsch *et al.*, 2011). For the resulting consensus samples, we argue that there is reasonable certainty regarding their subtypes. This enabled us to construct subtype predictors from consensus sets via supervised learning. For SSPs this may be especially advantageous as in this way a potentially unstable hierarchical clustering stage (Kapp *et al.*, 2006; Pusztai *et al.*, 2006; Haibe-Kains *et al.*, 2012) in the predictor construction phase can be completely avoided.
We start with a recapitulation of earlier findings and present a comprehensive reassessment of the concordance of the classic SSPs, including estimates based on subtype assignments taken from recent literature. We proceed with the construction of five consensus sets and construct a variety of CS-based models, which for a given predictor type (either SCM, SSP, or STG) mainly differ in the associated consensus training set and/or the gene list on which they were based. The CS-based predictors were subsequently applied to a large body of validation sets. In total, we collected 22 uniformly preprocessed datasets containing over 4,000 unique hybridizations. We used this compendium to assess the concordance of the classic SSPs and SCMs, and of nine CS-based subtype predictors: three SSPs, three SCMs, and three STGs.

3.3 Results

We investigated the ability to concordantly assign breast tumour tissues to the four main subtypes on which broad agreement exists (Guiu *et al.*, 2012), i.e. basal, HER2, luminal A and luminal B. Subtype assignments were based on three types of predictors: (*i*) SSPs, (*ii*) SCMs and (*iii*) STG subtype predictors derived from the gene expression-based quantification of estrogen receptor (ER), epidermal growth factor receptor 2 (HER2), progesterone receptor (PGR) and proliferation activity following the St. Gallen surrogate intrinsic subtype definitions (Figure 3.1). Based on these three subtyping schemes, five consensus sets were built for five different datasets (Figure 3.2A). On these a variety of consensus set-based SSP, SCM and STG predictors were constructed (Figure 3.2B), specifically designed to be concordant on the individual sample level, i.e. to show almost perfect levels of concordance on training data. These were subsequently applied to a large collection of validation sets. Our main research question is how concordant these predictors remain under varying conditions.

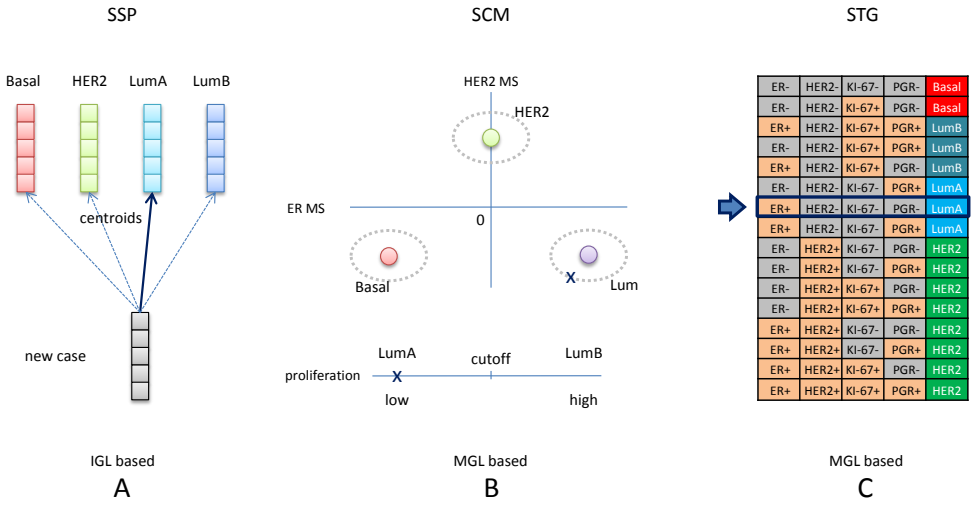


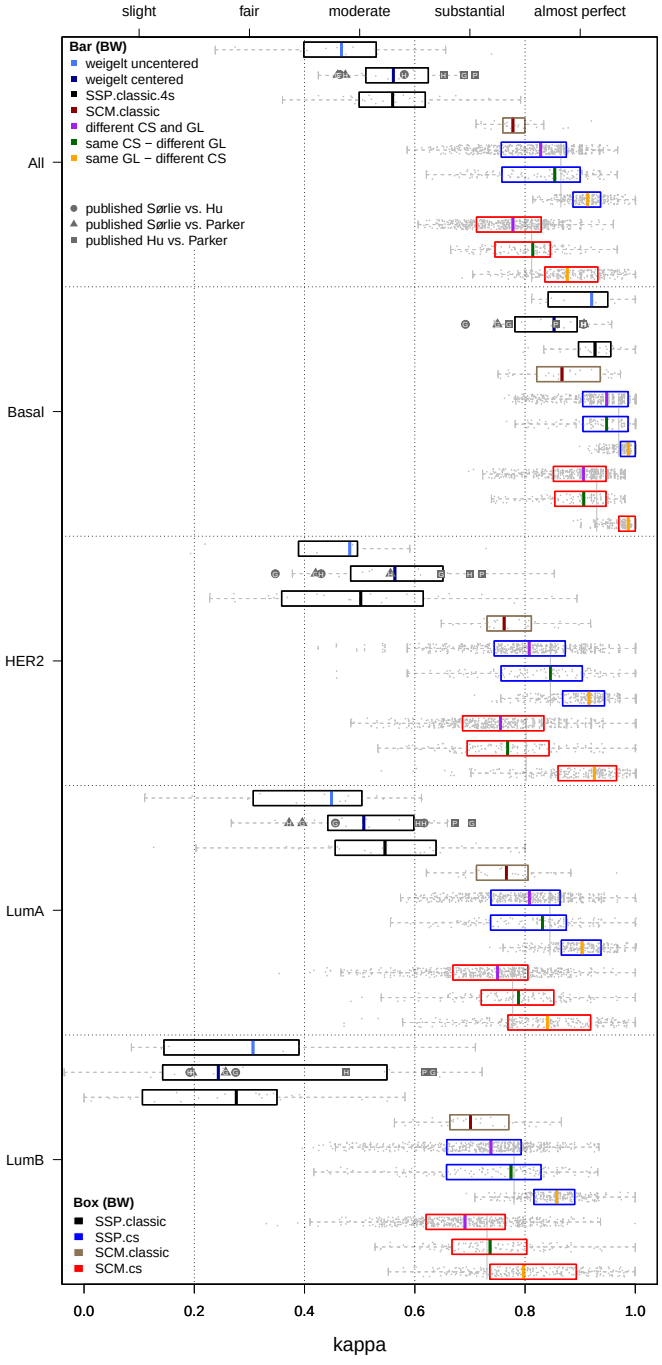
Figure 3.1: Conceptual overview predictors. A) Single sample predictor (SSP). For each subtype a centroid is computed (depicted by different colors) representing a vector of average values for each gene in the intrinsic gene list (IGL), i.e. a predetermined list of relevant genes, taken over a training set of samples assumed to be of the same subtype. In order to determine the subtype of a new case, one computes the distance to each of the centroids and assigns the new case to the subtype corresponding to the centroid that is nearest, here assumed to be the luminal A centroid, leading to the luminal A subtype. **B**) Subtype classification model (SCM). Each sample is represented by three module scores (MS) calculated based on module gene lists (MGLs), i.e. the list of genes associated with a module. Training set samples are first divided into basal, HER2 and luminal subtypes by fitting a 3-component Gaussian mixture model to the ER and HER2 related module scores (top panel, colored circles and dotted grey ovals). Subsequently, cases of the luminal subtype are divided into two subtypes, based on their proliferation module score. Samples with a low proliferation score are assigned to the lumA (luminal A) subtype, whereas samples with a high proliferation score are assigned to the lumB (luminal B) subtype. The subtype of a new case can be determined by calculating the posterior membership probabilities under the Gaussian mixture model and selecting the subtype associated with the maximum posterior probability. In the example, the new case (depicted with a cross) has a high ER module score and low HER2 and proliferation module scores, leading to the luminal A subtype. **C**) STG subtype predictor based on the St. Gallen surrogate intrinsic subtype definitions (Goldhirsch *et al.*, 2011). Over(+)/under(-)expression of clinical markers for ER, HER2, KI-67 (proliferation status) and PGR allows for $2^4 = 16$ distinct profiles. Here, the over/underexpression status of each marker was determined based on microarray measurements in a way similar to SCMs, i.e. via module scores. The subtype of a new case is fully determined by the over/underexpression status of the individual markers. In the example, the new case is assumed to have a high ER signaling score and low HER2, PGR and proliferation scores, leading to the luminal A subtype (blue arrow). A more comprehensive description of each subtype predictor type is provided in the supplementary information at the end of this Chapter.

Concordance assessments were made for all predictor pairs including the classic SSP and SCM predictors. The main results are captured in Figure 3.3 that presents the central figure of this text. The remainder of this section can be divided into two parts: (*i*) a recapitulation of previous findings, based on reported classic SSP subtype assignments, (*ii*) the construction of the consensus sets followed by evaluation of CS-based subtype predictors (see Materials and Methods) and their classic counterparts via intra- and inter-predictor concordance assessments.

Figure 3.2 ***(facing page)*****: Consensus set construction and evaluation scheme.** In panels (**A**) and (**B**), row (i) represents subtype predictor models and row (ii) corresponds to sets of predictions made by these models. Each column corresponds to a different predictor/subtype assignment set pair. **A**) Consensus set construction (see also Materials and Methods). For a given training set T, five initial sets of subtype assignments are obtained. First, the original PAM50 predictor is applied to T, resulting in subtype assignment set ST PAM. Next, three SCMs models are estimated on T, based on the MGLs D, W and HK (Supplementary Information). Here SCM X,T denotes an SCM estimated on T, based on MGL X. The resulting SCMs are subsequently applied to T resulting in three additional sets of subtype assignments, i.e. ST D, ST W, and ST HK. A final set of subtype assignments ST STG is obtained by the application of the STG predictor on the over/underexpression profile of ER, HER2, PGR and proliferation phenotypes, estimated on T. From the five subtype assignment sets in row (ii) a consensus set (CS) is derived consisting of those samples in T for which all five subtype assignments are concordant. **B**) Construction and evaluation of the consensus set-based subtype predictors SSP.cs (left), SCM.cs (middle) and STG.cs (right), see also Materials and Methods. For a given CS, three SSPs are constructed that differ only by their associated IGL S, H or P (Supplementary Information). Here SSP X,CS represents an SSP with associated IGL X, of which the centroids are estimated on CS. The SSPs are subsequently applied to a validation set V, leading to subtype assignment sets SV S, SV H and SV P, respectively. On the same CS also three SCMs are constructed, based on the MGLs D, W and HK. The resulting SCMs are subsequently applied to validation set V, yielding subtype assignment sets SV D, SV W and SV HK. Similar to SCMs, also three STG.cs predictors are constructed based on MGLs D, W and HK and applied to validation set V.

Gene list
Subtype predictor
Dataset
Subtypes
i
ii
1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
PAM 50
MGL D
MGL W
MGL HK
SCM D,T
SCM W,T
SCM HK,T
STG
T
ST PAM
ST D
ST W
ST HK
ST STG
CS
A
IGL S
IGL H
IGL P
MGL D
MGL W
MGL HK
SSP S,CS
SSP H,CS
SSP P,CS
SCM D,CS
SCM W,CS
SCM HK,CS
STG D,CS
STG W,CS
STG HK,CS
V
SV S
SV H
SV P
SV D
SV W
SV HK
SV D
SV D
SSP.cs
SCM.cs
STG.cs
B

Figure 3.3 *(facing page)*: Intra-predictor concordance of SSPs and SCMs hgu133plus2 compendium. The five panels show box and whisker (BW) plots for kappa statistics calculated over all subtypes and for each subtype separately, as indicated on the left hand side. Results for individual datasets are superimposed as dots. Each panel contains ten BW plots. From top to bottom these respectively indicate concordance for pairs of: (*i*) classic SSPs initially reported by Weigelt *et al.* (2010a), i.e. based on uncentered data ('weigelt uncentered'), (*ii*) classic SSPs by Weigelt *et al.* (2010c), based on centered data ('weigelt centered'). All subtype assignments related to the BWs 'weigelt uncentered' and 'weigelt centered' were retrieved from `http://rock.icr.ac.uk/collaborations/Mackay/centroid.correlations.Eset/`. Estimates based on subtype assignments from recent literature are superimposed as gray symbols with letters (see running text), (*iii*) classic SSPs without a normal-like subtype (SSP.classic.4s), (*iv*) classic SCMs (SCM.classic), (*v*) SSP.cs predictors, different CS and IGL, (*vi*) SSP.cs, same CS and different IGL, (*vii*) SSP.cs, same IGL and different CS, (*viii*) SCM.cs, different CS and MGL, (*ix*) SCM.cs, same CS and different MGL, and (*x*) SCM.cs, same MGL and different CS. Results for BWs (*iii*)-(*x*) are based on the hgu133plus2 compendium. Vertical gray lines indicate kappa estimates that were pooled over all three groups of comparisons per predictor type. Top legend: type of concordance assessment indicated by the color of BW median values (indicated by a bar) (GL: gene list, IGL or MGL). Bottom legend: predictor type indicated by the color of a BW box. Numerical details of the BW plots are presented in Supplementary Tables 3.2 and 3.3.


slight
fair
moderate
substantial
almost perfect
Bar (BW)
weigelt uncentered
weigelt centered
SSP.classic.4s
SCM.classic
different CS and GL
same CS − different GL
same GL − different CS
published Sørlie vs. Hu
published Sørlie vs. Parker
published Hu vs. Parker
All
Basal
HER2
LumA
LumB
Box (BW)
SSP.classic
SSP.cs
SCM.classic
SCM.cs
0.0
0.2
0.4
0.6
0.8
1.0
kappa

3.3.1 Concordance of classic SSPs

Initial findings Weigelt et al. and subsequent reactions

The initial study by Weigelt *et al.* (2010a) reported low concordance levels for the classic SSPs. Those findings were based on data from four datasets, profiled on different array platforms, with a total of 832 samples. For each dataset separately the level of concordance between subtype assignments (including the normal-like subtype) as obtained by different SSP pairs was measured using Cohen's kappa statistic. The top box and whisker (BW) plot in each panel of Figure 3.3 ('weigelt uncentered') shows the concordance levels calculated based on the subtype assignments reported by Weigelt *et al.* (2010a) (normal-like not shown). The classic SSPs were overall only moderately concordant (median κ=0.467, Supplementary Table 3.2). Only the basal subtype could be assigned with almost perfect levels of concordance (median κ=0.921). The study, however, was strongly criticized in subsequent reactions by Perou *et al.* (2010) and Sørlie *et al.* (2010) as their results were based on uncentered data. The lack of data centering is especially problematic when analyzing single-channel datasets (Supplementary Table 3.4). Unfortunately, three of the four studies selected by Weigelt *et al.* (2010a) were single-channel datasets. A lack of data centering in such cases leads to a strong increase of the correlation to the luminal B centroid (Sørlie *et al.*, 2010). Furthermore, correlations to the normal-like group are strongly decreased, in some cases to such an extent that the normal-like subtype can no longer be detected (Supplementary Figure 3.1). In a reanalysis of their own data, Weigelt and colleagues (Weigelt *et al.*, 2010c) showed that centering, however, did not lead to a substantial improvement in terms of concordance. Concordance levels calculated based on the subtype assignments reported by Weigelt *et al.* (2010c) are depicted by a second set of BW plots in Figure 3.3 ('weigelt centered', median κ=0.561). Our reanalysis shows that for single-channel datasets, the effect of centering or not is in fact as large as the effect of a change to another SSP as studied by Weigelt and colleagues (Supplementary Table 3.4). From the latter observation the criticisms expressed by Perou *et al.* (2010) and Sørlie *et al.* (2010) appear justified.

Confirmation lack of concordance of classic SSPs on large sample sizes

The results above were based on only four datasets of limited size. In order to increase the sample size, we compiled a large set of reported subtype assignments based on the efforts of three research groups. The main results are superimposed over the 'weigelt centered' BW plots in Figure 3.3 as gray

Dataset	Chip	Nr. of samples after QC	Nr. of samples (%)				
			CS	Basal	HER2	LumA	LumB
Bos	hgu133plus2	188	119 (63.3)	49 (41.2)	19 (16.0)	23 (19.3)	28 (23.5)
expO	hgu133plus2	333	213 (64.0)	56 (26.3)	20 (9.4)	75 (35.2)	62 (29.1)
Guedj	hgu133plus2	501	235 (46.9)	40 (17.0)	21 (8.9)	88 (37.4)	86 (36.6)
Li	hgu133plus2	109	83 (76.1)	25 (30.1)	10 (12.0)	29 (34.9)	19 (22.9)
Sabatier	hgu133plus2	242	162 (66.9)	63 (38.9)	15 (9.3)	40 (24.7)	44 (27.2)
Total		1373	812 (59.1)	233 (28.7)	85 (10.5)	255 (31.4)	239 (29.4)

Table 3.1: Overview of the five training sets used for consensus set construction and the resulting consensus sets. See also Supplementary Table 3.1. Numbers in parentheses represent percentages. For CS, percentages were calculated w.r.t. the number of samples after QC; for the subtypes w.r.t. the size of the CS.

symbols. Each symbol indicates a particular pair of classic SSPs (see legend), while letters indicate the origin of the subtype assignments, i.e. G: (Guedj *et al.*, 2011), H: (Haibe-Kains *et al.*, 2012) and P: Perou lab (`https://genome.unc.edu/pubsup/breastGEO/`). These findings clearly confirm the main claim by Weigelt et al. namely the lack of concordance of the classic SSPs, on a much larger number of samples. Especially the luminal B subtype was highly discordant (κ=0.192-0.633, Supplementary Table 3.5). In agreement with previous observations the basal subtype was most concordantly subtyped (κ=0.692-0.907). The highest level of overall concordance between SSPs was obtained by the Perou lab for the SSP by Hu and PAM50 (κ=0.710, cc=77.60%). This is not surprising given that both SSPs were developed at the Perou lab and were mainly applied by them to data from the same dual-channel platform.

3.3.2 Consensus set construction and predictor evaluation

Table 3.1 presents an overview of the five training sets used for consensus set construction (Figure 3.2A), as well as a decomposition by subtype for each of the resulting consensus sets. For every CS all four subtypes, i.e. basal, HER2, luminal A and luminal B, were well represented. The stringent CS selection criteria implied a strong reduction in terms of samples available for predictor construction (median 64.0% remaining). On each consensus set three SSP, three SCM and three STG predictor models were constructed via supervised learning (Figure 3.2B). Consensus set samples have a number of desirable properties. They can be stably identified using hierarchical clustering, lead to module scores that are reasonably bimodal and - most importantly - consensus set-based predictors concordantly subtype each others samples (see Supplementary Information). Hence, CS-based predictors are highly concordant on the individual sample level on training data.

We next investigated how concordant the CS-based predictors remained when

they were evaluated on a large set of uniformly preprocessed validation datasets measured on Affymetrix hgu133plus2 and hgu133a microarrays (Supplementary Table 3.1). In order to further minimize differences due to technical variation, in the main text we mainly present results on the hgu133plus2 compendium, consisting of 11 datasets (2,019 samples after QC, Supplementary Table 3.1). Results for the analyses on all 3,908 arrays are reported in the Supplementary Figures and Tables and showed highly similar results. Based on the subtype assignments on the validation sets, different types of comparisons can be made. We distinguish differences in subtype assignments due to differences in: (*i*) the consensus training set used for predictor construction, (*ii*) the selected IGL or MGL and (*iii*) the predictor type. We considered two types of predictor comparison, i.e. intra-predictor and inter-predictor type comparisons. Intra-predictor comparisons only involve comparisons between predictors of the same type, e.g. the SSP of Hu vs. the SSP of Parker. Inter-predictor comparisons only involve comparisons between predictors of different types, e.g. SCM vs. SSP.

Classic SSP intra-predictor evaluation with and without a normal-like subtype

The classic SSP concordance estimates presented above were based on previously reported subtype assignments that included a normal-like subtype. We also estimated these on our hgu133plus2 compendium and again only moderate levels of agreement between classic SSPs were observed (median κ=0.575, median cc=70.75%; Supplementary Table 3.3). SCM predictors, as well as our CS-based predictors, however, do not consider a normal-like subtype. The primary motivation for this choice is that currently there is no consensus whether this subtype is a genuine breast cancer subtype (Guedj *et al.*, 2011) or an artifact of breast tumour tissues having a high percentage of normal contamination in the tumour specimen (Parker *et al.*, 2009). Although the PAM50 predictor does include a normal-like subtype, this classification is merely considered as a quality-control measure (Parker *et al.*, 2009). In the remainder we do no longer consider the normal-like subtype and focus on the identification of the remaining subtypes instead. The third BW plot in each panel of Figure 3.3 (SSP.classic.4s, where '4s' indicates that we consider four subtypes instead of five) shows the concordance of the classic SSPs on our hgu133plus2 compendium when the normal-like centroid is removed. In this scenario we obtained similar kappa statistics for the classic SSPs as above (median κ=0.560, median cc=66.97%; Supplementary Table 3.3).

Classic SCM intra-predictor evaluation

In our compendium the concordance of the classic SCMs was substantially higher than for the classic SSPs and in the upper range of substantial agreement (median $\kappa = 0.778$, median cc=83.88%; Figure 3.3, Supplementary Table 3.3). Lowest concordance was observed for the luminal B subtype (median κ=0.701). Kappa statistics here are higher than those reported in Haibe-Kains *et al.* (2012), Table 3, where concordance between the three classic SCMs reached an average κ=0.720 (median κ=0.700). In our case, however, the classic SCMs were all constructed and evaluated using data measured on a single array design, whereas Haibe-Kains et al. constructed the classic SCMs on Affymetrix data and evaluated them on a compendium that also contained many non-Affymetrix datasets. When excluding the non-Affymetrix datasets, the concordance estimates for the classic SCMs based on the subtype assignments reported by Haibe-Kains *et al.* (2012) are highly similar to ours (Supplementary Table 3.6).

Strong increase in intra-predictor concordance for CS-based SSPs

The concordance levels of the consensus set-based SSPs, denoted as SSP.cs, showed a vast improvement w.r.t. the classic SSPs with kappa statistics in the range of almost perfect agreement (median κ=0.865, median cc=90.32%; Supplementary Table 3.3). Note that 5 of the 11 hgu133plus2 validation sets were also used for the construction of the consensus sets and CS-based predictors. In order to avoid an upward bias of the concordance of CS-based predictors, the reported kappa statistics are strictly based on those combinations where the training set and the validation set were different. Subtype-specific performances were equally strong with median kappa statistics of 0.970, 0.846, 0.845 and 0.780 for the subtypes basal, HER2, luminal A and luminal B, respectively. In order to investigate differences due to a change in IGL or consensus set in more detail, kappa statistics were partitioned into three disjoint groups (blue BW plots in Figure 3.3) for SSPs in which (*i*) both the consensus set and IGL were different, (*ii*) only the IGL was different and (*iii*) only the consensus set was different. As expected, concordance was lowest when both elements were different (median κ=0.828, Supplementary Table 3.2). Surprisingly, the impact of changing the IGL was larger than of a change to another consensus set (median κ=0.854 vs. κ=0.914). Consistent with previous literature, the luminal B subtype was most susceptible to changes in both the consensus set and IGL (median κ=0.738). However, when only the consensus set was changed, consensus for luminal B was still in the range of almost perfect agreement (median κ=0.857).

SCM.cs intra-predictor concordance

SCM predictors trained on consensus sets (SCM.cs) were also strongly concordant (median κ=0.812, median cc=86.67%; red BW plots in Figure 3.3, Supplementary Table 3.3), however, notably less than the SSP.cs predictors. The change to another MGL as compared to a change of consensus set showed a substantial loss in agreement (median κ=0.814 vs. κ=0.876). When both elements were changed, concordance dropped to the range of substantial agreement (median κ=0.778), a value equal to the overall concordance observed for the classic SCMs. Hence, SSP predictors benefit more from the consensus set construction scheme than SCMs.

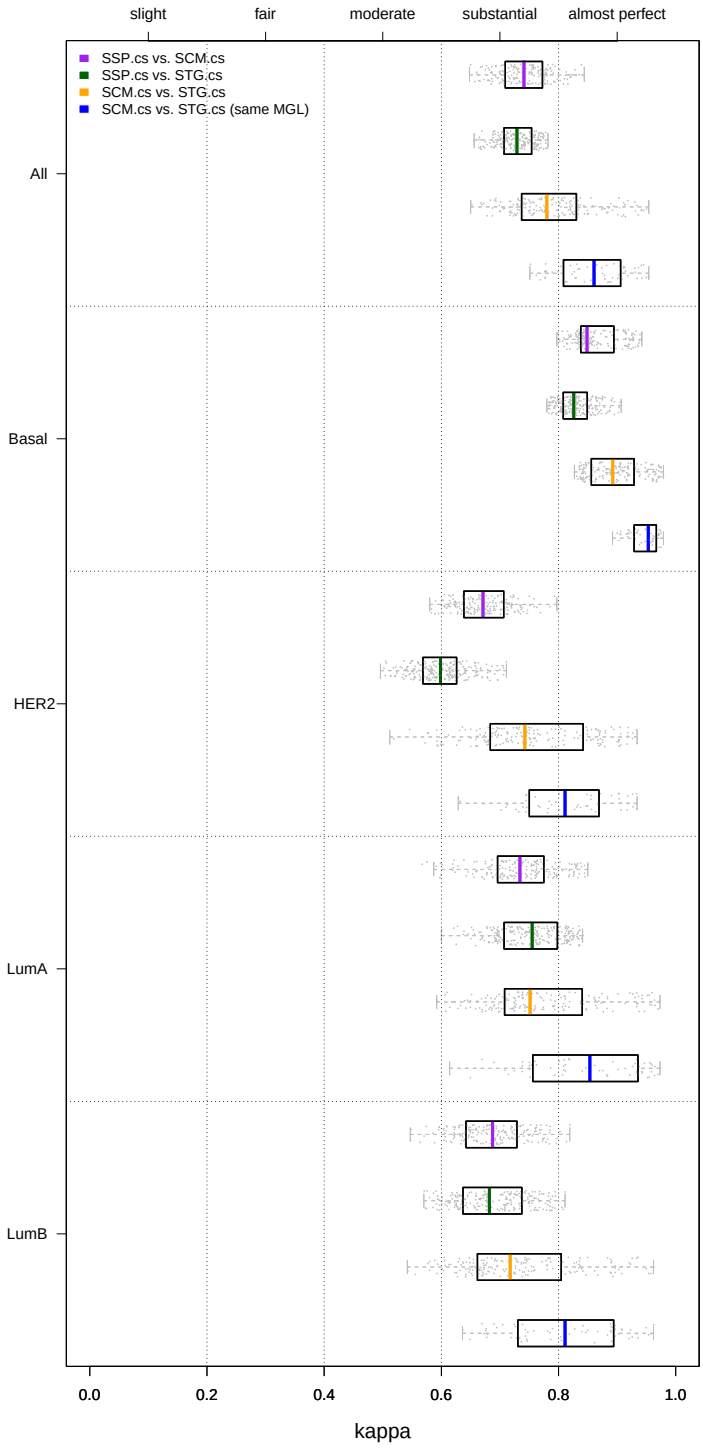
Concordance of CS-based models and their classic counterparts

When based on the same MGL the SCM.cs predictors showed almost perfect levels of concordance with their classic counterparts (median κ=0.893-0.926, median cc=92.15-94.55%; Supplementary Figure 3.2, Supplementary Table 3.7), with equally strong subtype-specific levels of agreement. A similarly strong level of concordance was observed between the classic PAM50 predictor and its CS-based counterpart based on IGL P (median κ=0.870, median cc=90.77%). For the two oldest SSPs by Hu and Sørlie, however, only substantial (SSP Hu: median κ=0.775, median cc=83.95%) and moderate (SSP Sørlie: median κ=0.584, median cc=70.24%) levels of concordance were obtained with their CS-based counterparts, respectively. Note that the high concordance of CS-based models with the classic SCMs and PAM50 implies that they share the strong prognostic value reported for classic subtype predictors (Parker *et al.*, 2009; Haibe-Kains *et al.*, 2012).

Inter-predictor concordance of CS-based SSPs and SCMs is only substantial

Weigelt *et al.* (2010a,c) mainly consider SSP intra-predictor concordance, i.e. concordance between predictors of the same type. Above, we showed that the

Figure 3.4 ***(facing page)*****: Inter-predictor concordance assessment.** The five panels show box and whisker plots for kappa statistics calculated over all subtypes and for each each subtype separately, as indicated on the left hand side. Results for individual datasets are superimposed as dots. The upper three BW plots in each panel show the inter-predictor concordance estimates between the SSP.cs, SCM.cs and STG.cs predictors pairs, as indicated by the legend. The bottom BW plot in each panel provides the concordance estimates for SCM.cs and STG.cs predictor pairs when based on the same modules, i.e. MGLs (with exception of PGR). Results are based on the hgu133plus2 compendium. Numerical details of the BW plots are presented in Supplementary Table 3.8.


slight
fair
moderate
substantial
almost perfect
SSP.cs vs. SCM.cs
SSP.cs vs. STG.cs
SCM.cs vs. STG.cs
SCM.cs vs. STG.cs (same MGL)
All
Basal
HER2
LumA
LumB
0.0
0.2
0.4
0.6
0.8
1.0
kappa

intra-predictor concordances for CS-based SSPs and SCMs are in the range of almost perfect concordance. In the challenging scenario in which the consensus training set, predictor type and (as a consequence) the gene list, are different we observed only substantial levels of concordance when comparing SSP.cs and SCM.cs predictors (median κ=0.741; median cc=81.02%; Figure 3.4, Supplementary Table 3.8), despite the fact that the CS-based predictors showed almost perfect levels of concordance on the consensus sets themselves. In line with previous observations, only the basal subtype was identified with almost perfect levels of agreement (median κ=0.849), while the luminal B and HER2 subtype assignments were least concordant (median κ=0.688 and κ=0.671, respectively).

High inter-predictor concordance of CS-based SCMs and STGs

So far we mainly focused on SSP and SCM-based approaches. We now consider in more detail the third subtype predictor type (STG; see Figures 3.1 and 3.2B), based on the St. Gallen surrogate intrinsic subtype definitions (Goldhirsch *et al.*, 2011). When based on the same MGL, SCM.cs and STG.cs models show almost perfect concordance (median κ=0.861; median cc=89.84%; Figure 3.4, Supplementary Table 3.8). The SSP framework is conceptually quite different and overall concordance between STG.cs and SSP.cs models is indeed considerably lower (κ=0.729). Interestingly, the lowest concordance between STG.cs and SSP.cs models was not obtained for the luminal B subtype, but for the HER2 subtype (median κ=0.599). Note that even though the STG.cs predictors represent only a simple rule-based subtyping prediction scheme, fully defined by the over/underexpression status of four markers, their intra-predictor concordance was the highest of all predictors considered when based on the same MGL (Supplementary Figure 3.5).

3.4 Discussion

A limitation of previous studies that assessed the concordance between subtype assignments (Weigelt *et al.*, 2010a; Perou *et al.*, 2010; Sørlie *et al.*, 2010; Weigelt *et al.*, 2010c; Haibe-Kains *et al.*, 2012) is that subtype predictors were evaluated in what could be considered a worst-case scenario. Next to differences in gene lists, reported concordance statistics may have been negatively influenced by differences in the training sets used and technical heterogeneity, e.g. differences in microarray platforms, normalization and scaling strategies. Moreover, robustness and concordance of SSPs may have been negatively affected by the instability of the hierarchical clustering step (Tibshirani and Walther, 2005; Pusztai *et al.*, 2006; Lusa *et al.*, 2007; Mackay *et al.*, 2011). Our goal was to

design an experimental setup that disentangles the various factors influencing concordance estimates, in order to obtain an improved perspective on the behaviour of modern subtype predictor schemes such as PAM50 (Parker *et al.*, 2009) and SCMs (Desmedt *et al.*, 2008; Wirapati *et al.*, 2008; Haibe-Kains *et al.*, 2012).

3.4.1 Standardization of microarray data

In contrast to the studies by Weigelt *et al.* (2010a,c) and Haibe-Kains *et al.* (2012), we constructed and evaluated predictors on data from a single measurement platform only, i.e. Affymetrix. Previously reported subtype assignments provide some evidence of the detrimental impact of technical heterogeneity (Supplementary Table 3.6), suggesting a decrease in performance when evaluating predictors in a multi-platform setup. In our study, all arrays were treated identically via a three-step procedure which involved a stringent quality control stage, renormalization of the intensity data by frozen RMA (McCall *et al.*, 2010) and a subsequent robust scaling step. The quality of the resulting data was further supported by the high concordance obtained on replicate array pairs (Supplementary Figure 3.6, Supplementary Table 3.9). In this standardized setup, we observed only a slight decrease in concordance when evaluating the CS-based predictors on data from another array design (hgu133a) than the one on which they were constructed (hgu133plus2), see Supplementary Tables 3.2 and 3.3.

3.4.2 Importance of consensus set

In our setup, predictor construction was performed on carefully designed training sets. Only those samples were used of which the subtypes could be concordantly identified across multiple sources, i.e. the consensus set samples. The idea of a consensus set is reminiscent of the use of a core set of samples in most hierarchical clustering based subtyping approaches. From all clustered samples in general a selection is made in order to exclude samples with low correlation to each subtype. Core set selection is based on heuristics (Sørlie *et al.*, 2003; Hu *et al.*, 2006) or statistical methods that assess the stability of a hierarchical clustering (Alexe *et al.*, 2006; Parker *et al.*, 2009). Guedj *et al.* (2011) constructed a core set by selecting those samples that were assigned to the same subtype by three different clustering methods, viz. hierarchical clustering, k-means and Gaussian mixture models. In contrast to these approaches, our consensus set inclusion criteria are stricter and also incorporate differences in gene lists. Since there is reasonable certainty regarding the subtype classification of the consensus set samples, we hypothesized that subtype predictors can safely be constructed on a consensus set via supervised

learning. Indeed, our results show that the subtype classification of the consensus set samples themselves is highly concordant (median κ=0.957; Supplementary Information, Table 5). Another important advantage of using consensus sets for predictor construction is that SSPs, SCMs and STGs can be constructed on identical training sets. This allowed us to establish that the influence of a change in gene list is larger than that induced by a change in consensus training set. Changing both elements still led to (close to) almost perfect concordance (SSP.cs: median κ=0.828, SCM.cs: median κ=0.778). For SSPs our concordance estimates are considerably higher than those reported by Weigelt *et al.* (2010a,c) (median κ=0.467 before centering, median κ = 0.561 after centering) and Haibe-Kains *et al.* (2012)(κ=0.45-0.58). Concordance reported for the classic SCMs trained on the expO dataset (κ=0.65-0.81) (Haibe-Kains *et al.*, 2012) is also lower but more comparable to ours (SCM.cs, different MGL: median κ=0.814). If we consider only subtype assignments on Affymetrix cohorts, estimates on the concordance of the classic SCMs (Haibe-Kains *et al.*, 2012) (Supplementary Table 3.6) are highly similar to those reported here. SSPs appear to benefit more from the consensus set approach than SCMs. This is likely due to the fact that in our setup no hierarchical clustering stage was required in order to construct SSPs. For SCMs it may actually not be necessary to identify a consensus set for model fitting purposes. We observed almost perfect levels of concordance between SCM models based on consensus set samples only and those fitted on complete cohorts (median κ=0.954; median cc=96.67%). In this respect SCMs are clearly superior in terms of robustness compared to SSPs constructed via hierarchical clustering.

3.4.3 Factors influencing concordance

Prat *et al.* (2011) recommend the highest level of concordance, i.e. almost perfect concordance for routine clinical use of pathology and gene-expression-based tests. Their comprehensive review shows that for virtually all currently used biomarkers in breast cancer only substantial or moderate concordance between two different methods has been reported. They claim that almost perfect concordance can only be achieved by using a single platform and a standardized protocol for such tests. Our experimental setup provides an improved perspective on the factors influencing concordance between different subtyping schemes. When comparing different SSPs trained on different consensus sets, we moved from moderate concordance (Weigelt *et al.*, 2010a; Haibe-Kains *et al.*, 2012) to almost perfect concordance. These results clearly illustrate the large benefit of using a standardized approach. The inter-predictor results, however, show that the choice of predictor type and associated gene lists matters. We observed large differences in the subtype

assignments from predictors of different types. In the most challenging scenario in which training set, predictor type and gene list are different, we moved from moderate concordance (median κ=0.5) (Haibe-Kains *et al.*, 2012) to substantial concordance (median κ=0.741; Supplementary Table 3.8). Even though we based our conclusions on research data, we feel such discrepancies are an impediment to their incorporation into clinical practice as it is clear that the specific choice of a predictor type matters, yet it is unclear which predictor type is to be preferred. In the scenario analysed by Weigelt *et al.* (2010d) one could argue that the PAM50 predictor presents an evolution over time in which deliberate design changes were made with respect to older SSPs (Perou *et al.*, 2010) and one may therefore claim that the observed discordance is a feature instead of a flaw. In the scenario analysed here, however, there is little room for such an interpretation as all predictors were specifically designed to be concordant on the individual sample level, while the influence of technical heterogeneity was strongly reduced. Our results also show large differences in concordance for the different subtypes. In general, the basal subtype was the only subtype which could consistently be identified with almost perfect concordance (Supplementary Table 3.2), as reported previously (Weigelt *et al.*, 2010a; Haibe-Kains *et al.*, 2012).

The observed intra- and inter-predictor discordances can be explained by various factors. Our experiments clearly highlight the importance of the selected gene list, whose influence was consistently larger than the choice for a particular training set during predictor construction. Of the intrinsic subtypes the luminal B subtype was the most challenging subtype to detect concordantly. When based on the same gene list, however, we still obtained concordance levels in (or extremely close to) the range of almost perfect agreement (SSP.cs: median κ=0.857, SCM.cs: median κ=0.797, Supplementary Table 3.2). To a certain degree, discordance between luminal A and luminal B subtype assignments may be expected if proliferation indeed forms a continuum, as suggested before (Weigelt *et al.*, 2010d; Haibe-Kains *et al.*, 2012). In most datasets considered here, however, the proliferation markers were bimodal, albeit almost never strongly (Supplementary Table 3.10). The observed lack of inter-predictor concordance can be further explained by differences in model assumptions and subtype definitions. Note that, after more than a decade of molecular breast cancer subtyping, there still is no consensus on both the number and definitions of breast cancer subtypes. Especially problematic is the relation of HER2 to the other subtypes. HER2 has often been considered to belong to the ER- branch of subtypes, as is the case for the original St. Gallen surrogate intrinsic subtype definitions consisting of 5 subtypes (Goldhirsch *et al.*, 2011). In these the luminal B subtype is split into two subtypes, i.e. luminal B (HER2+) and luminal B (HER2-) (Supplementary Figure 3.7A). In

order to obtain a 4-subtype taxonomy as considered in this chapter, we mapped the luminal B/HER2+ subtype to the HER2 subtype and luminal B/HER2- to the luminal B subtype. This mapping was chosen as it maximizes similarity with SCMs, in which HER2 subtype assignments are possible for both ER- and ER+ samples (Desmedt *et al.*, 2008) (Supplementary Figure 3.7B). This mapping, likely, has a positive effect on the inter-predictor concordance of STG.cs and SCM.cs predictors. However, discordance may still arise between SCMs and STGs due to the PGR status, which is not considered by SCMs. Finally, we note that various recent studies have shown that within each of the intrinsic subtypes there still is considerable heterogeneity left (Lehmann *et al.*, 2011; Cancer Genome Atlas Network, 2012; Curtis *et al.*, 2012). Molecular heterogeneity within a subtype does not imply discordance as studied in this chapter per se. However, this changes when it affects more than one of the intrinsic subtypes, as is the case in the St. Gallen criteria. Therefore, in future concordance studies it is likely that considerable discordance will remain to be observed until the definitions of the molecular subtypes have been sufficiently refined.

In conclusion, we presented a comprehensive evaluation of SSP and SCM subtype predictors instigated by the Lancet Oncology article by Weigelt *et al.* (2010a) and subsequent reactions (Perou *et al.*, 2010; Sørlie *et al.*, 2010; Weigelt *et al.*, 2010c). The initial study by Weigelt and colleagues reported low concordance between subtype assignments based on the classic SSPs and concluded that SSPs do not reliably assign subtypes to individual patients. In contrast, our findings show that in a carefully standardized setup via the use of consensus sets almost perfect concordance can be achieved by both SSP and SCM predictor types and for multiple gene lists. However, differences between predictor types, gene lists and training datasets combined, result in subtype assignments that only show substantial levels of agreement. Prospective clinical trials are needed to go beyond the concordance issues investigated in this chapter and to determine which subtype predictor is most relevant for predicting treatment response of an individual patient.

3.5 Materials and Methods

3.5.1 Gene expression data

A breast cancer compendium consisting of 22 datasets was constructed. The compendium comprises 4,227 breast cancer tumour samples (Supplementary Table 3.1) and includes a set of 93 replicate array pairs. All datasets were obtained using a single measurement platform, i.e. Affymetrix. Each of the hybridizations was uniformly processed by a three-step procedure consisting of

(*i*) re-normalization by frozen RMA (fRMA) (McCall *et al.*, 2010), (*ii*) quality control and (*iii*) a robust scaling step, as described below.

fRMA normalization Our Affymetrix compendium was normalized by a modified version of the RMA methodology, i.e. frozen RMA. An important distinction between default RMA and fRMA is that in fRMA the reference distribution is not estimated on each dataset separately, but a pre-computed, i.e. frozen, reference distribution is employed. Here, a single reference distribution was used for all 22,215 non-control probesets present on the hgu133a platform. Expression estimates were based on the robust weighted average mode (McCall *et al.*, 2010) of fRMA. A more complete description of the normalization process is provided in the Supplementary Information.

Quality control An extensive quality control (QC) analysis was performed aimed at identifying hybridizations that consistently showed indications of poor quality, either before or after normalization. The complete QC protocol, including related results, is described in the Supplementary Information. In total 319 samples (7.55%) were rejected based on consistent indications of poor quality. In the remaining analyses only hybridizations that passed QC were used.

Robust scaling The datasets comprising the Affymetrix compendium were compiled over a large number of years and involve a substantial number of distinct processing sites, i.e. research institutes. Therefore, it is likely that there are substantial batch effects between datasets (Scherer, 2009; Leek *et al.*, 2010). Although normalization by fRMA is especially useful in such a scenario, it does not completely remove all batch effects. Therefore, for SSP-related experiments after normalization by fRMA the expression of each of the datasets D1-D22 (Supplementary Table 3.1) was robustly scaled (Haibe-Kains *et al.*, 2012), using the *genefu* package. In the scaling step, for each dataset and probeset separately, the 2.5 and 97.5 percentiles were scaled to -1 and +1, respectively. For a given SCM or STG and dataset, instead of scaling the expression data directly, we first computed the module scores on unscaled data and subsequently robustly scaled the module scores.

3.5.2 Consensus sets and CS-based predictor construction and evaluation

In order to obtain predictors that are as concordant as possible on the individual sample level, for a given training set T, we only used those samples for predictor construction that were concordantly subtyped across five distinct predictors:

(*i*) the PAM50 SSP, (*ii*) three SCMs estimated on T, based on the MGLs D, W and HK and (*iii*) an STG predictor estimated on T (Supplementary Information). We refer to the set of concordantly subtyped samples as the consensus set (CS) of T. The complete procedure is outlined in Figure 3.2A. Note that of the five predictors used to determine a CS, four are constructed via unsupervised learning on T itself. An advantage of using consensus sets for predictor construction is that SSPs, SCMs and STGs can be constructed on identical sample cohorts. Furthermore, SSPs can be constructed in a supervised way, i.e. a potentially highly unstable hierarchical clustering step (Haibe-Kains *et al.*, 2012) can be avoided completely.

Construction of CS-based models For any given CS, three SSPs, three SCMs and three STGs were constructed using the consensus training sets. For SSP construction we employed the IGLs related to the classic SSPs, i.e. IGL S, H and P (Supplementary Information), and used the updated probeset-to-gene mappings of Mackay *et al.* (2011). Similarly, for SCMs we used the MGLs related to the classic SCMs, i.e. the MGL D, W and HK. For all IGLs and MGLs, in case multiple probesets mapped to the same Entrez Gene ID, the most variable probeset was selected (Haibe-Kains *et al.*, 2012). Note that SCMs consider three out of the four biological processes included in STGs, i.e. ER and HER2 signaling and proliferation. Given the intrinsic similarity between SCMs and STGs, it is interesting to study them in a scenario in which common markers are tracked by identical modules. We therefore constructed a variety of CS-based STGs in which ER, HER2 and proliferation phenotypes were measured by the same modules as for SCMs, i.e. MGLs HK, D and W. As SCMs do not consider PGR, for this marker we always used the same single probeset module (Supplementary Information). We refer to the resulting CS-based predictors as SSP.cs, SCM.cs and STG.cs predictors, respectively. After predictor construction, all predictors were applied to a large collection of validation sets, of which the resulting subtype assignments were subsequently used in various concordance assessments. The complete procedure is outlined in Figure 3.2B. For SCM.cs we used the *subtype.cluster* function in the Bioconductor package *genefu*, which for a given consensus training set and MGL computes the module scores and estimates the parameters of the associated mixture model.

Concordance measure The level of concordance between subtype assignments of two distinct subtype predictors was measured by the percentage of concordant samples (*cc*) and Cohen's kappa statistic (Cohen *et al.*, 1960). The range of values kappa can take is often subdivided into five intervals that describe concordance in qualitative terms: 0-0.2 (slight), 0.21-0.4 (fair),

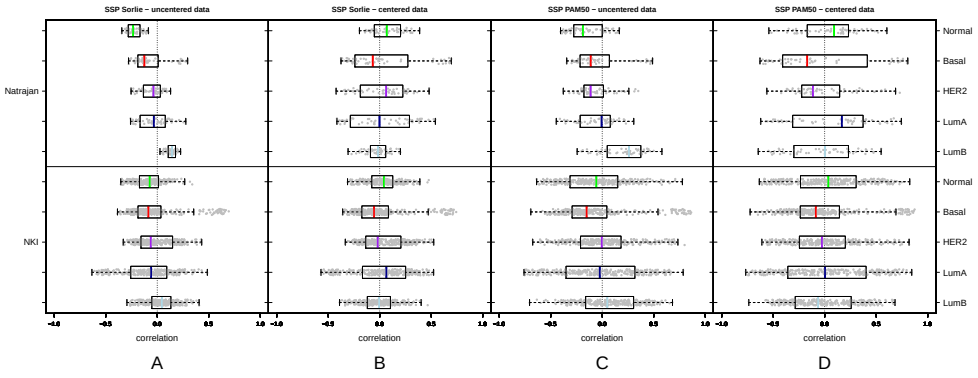
0.41-0.6 (moderate), 0.61-0.8 (substantial) and 0.81-1 (almost perfect). Kappa statistics were computed over all subtypes or for a specific subtype only. In the latter case, for a given subtype s, the complete subtype vector was transformed into a binary vector indicating whether the prediction was either s or not s. Subsequently, a traditional contingency table was formed for which a kappa statistic was computed representing the subtype-specific kappa for subtype s.

3.6 Supplementary Figures and Tables

Supplementary Table 3.1 ***(facing page)*****: Overview Affymetrix compendium**. The compendium consists of data from 22 datasets measured by a single measurement platform, i.e. Affymetrix. The expression data was measured on two distinct array designs, i.e. hgu133plus2 (top 11 datasets, 2,182 samples) and hgu133a (bottom 11 datasets, 2,045 samples). We only considered the 22,215 probesets that these designs have in common, which represent all non-control probesets present on the hgu133a platform. Shared probesets are based on an identical set of probes with identical probe sequences. Remaining heterogeneity on these datasets was further reduced using frozen RMA (McCall *et al.*, 2010) normalization and robust scaling (Haibe-Kains *et al.*, 2012) (Materials and Methods, main text). Furthermore, an extensive quality control (QC) analysis was performed aimed at identifying (and removing) hybridizations that consistently showed indications of poor quality (Materials and Methods, main text; Supplementary Information). *ID*: short dataset identifier; *Dataset*: dataset name; *Nr. of samples*: total number of available samples; *Rejected*: number of samples removed based on QC; *Passed*: total number of samples remaining after QC. In total 319 samples (7.55%) were rejected based on consistent indications of poor quality. *Chip*: array design used, i.e. hgu133plus2 or hgu133a; *Source*: the accession number under which the raw intensity data can be found at GEO (Edgar *et al.*, 2002). Dataset D10 is available at ArrayExpress (Brazma *et al.*, 2003) (accession number E-MTAB-365); *Reference*: reference to main study. The 344 sample VDX dataset (D17) consists of the combined expression data of the 286 sample dataset by Wang *et al.* (2005) and the 58 ER- sample dataset by Yu *et al.* (2007). Finally, note that the Symmans datasets (D11-D13) represent ER+ datasets. To prevent bias due to scaling Perou, datasets D12 and D13 were first concatenated to the VDX dataset and subsequently scaled as a single dataset, after which the VDX dataset was removed. Similarly, dataset D11 was combined with the expO dataset during scaling. A similar strategy was followed by Haibe-Kains *et al.* (2012).

ID	Dataset	Nr. of samples	Nr. of samples (QC)		Chip	Source	Reference
			Rejected	Passed			
D1	Richardson (I)	47	5	42	hgu133plus2	GSE3744	Richardson *et al.* (2006)
D2	Li	115	6	109	hgu133plus2	GSE19615	Li *et al.* (2010)
D3	Lu	127	3	124	hgu133plus2	GSE5460	Lu *et al.* (2008)
D4	Bos	204	16	188	hgu133plus2	GSE12276	Bos *et al.* (2009)
D5	Dedeurwaerder	90	7	83	hgu133plus2	GSE20711	Dedeurwaerder *et al.* (2011)
D6	expO	353	20	333	hgu133plus2	GSE2109	Haibe-Kains *et al.* (2012)
D7	Kao	327	33	294	hgu133plus2	GSE20685	Kao *et al.* (2011)
D8	Richardson (II)	84	9	75	hgu133plus2	GSE18864	Li *et al.* (2010)
D9	Sabatier	266	24	242	hgu133plus2	GSE21653	Sabatier *et al.* (2011)
D10	Guedj	537	36	501	hgu133plus2	E-MTAB-365	Guedj *et al.* (2011)
D11	Symmans (III)	32	4	28	hgu133plus2	GSE17700	Symmans *et al.* (2010)
D12	Symmans (I)	298	23	275	hgu133a	GSE17705	Symmans *et al.* (2010)
D13	Symmans (II)	32	3	29	hgu133a	GSE17700	Symmans *et al.* (2010)
D14	Desmedt	198	13	185	hgu133a	GSE7390	Desmedt *et al.* (2007)
D15	Farmer	49	3	46	hgu133a	GSE1561	Farmer *et al.* (2005)
D16	Schmidt	200	18	182	hgu133a	GSE11121	Schmidt *et al.* (2008)
D17	VDX	344	29	315	hgu133a	GSE2034,GSE5327	Wang *et al.* (2005); Yu *et al.* (2007)
D18	Miller	251	18	233	hgu133a	GSE3494	Miller *et al.* (2005)
D19	Pawitan	159	16	143	hgu133a	GSE1456	Pawitan *et al.* (2005)
D20	Shi	278	19	259	hgu133a	GSE20194	Shi *et al.* (2010); Popovici *et al.* (2010)
D21	MSK	99	8	91	hgu133a	GSE2603	Minn *et al.* (2005); Haibe-Kains *et al.* (2010)
D22	UNT	137	6	131	hgu133a	GSE2990	Sotiriou *et al.* (2006); Haibe-Kains *et al.* (2010)
Total		4227	319	3908			

Supplementary Table 3.2 *(facing page)*: Intra-predictor concordance of SSPs and SCMs (hgu133plus2 compendium). Numerical details of Figure 3.3 in the main text. *Panel*: panel indicator; *Index*: box and whisker plot index per panel, starting from the top; *Description*: predictor pair description; *Type*: subtype predictor type; κ *(Weigelt)*: median kappa statistics based on the published subtype assignments by Weigelt *et al.* (2010a,c); κ *(133plus2)*: median kappa statistics computed over all 2,019 Affymetrix hgu133plus2 arrays (Supplementary Table 3.1); κ *(all)*: median kappa statistics computed over all 3,908 samples in the Affymetrix compendium, i.e. including the hgu133a samples; $\Delta\kappa$: difference between the median kappa statistics computed on hgu133plus2 samples and on all samples. A negative value indicates a decrease in concordance when the hgu133a samples are included.



Supplementary Figure 3.1: Example of impact of uncentered data in analysis of Weigelt and colleagues. Correlation between individual samples and each of the five subtype centroids of two classic SSPs (Sørlie and PAM50) for two datasets (Natrajan and NKI) analyzed in Weigelt *et al.* (2010a). Correlations are shown for the SSP of Sørlie based on uncentered data (**A**), for the SSP of Sørlie based on centered data (**B**), for the PAM50 subtype predictor on uncentered data (**C**) and for PAM50 on centered data (**D**). The upper half (Natrajan) in each panel shows correlations for a single-channel dataset (Natrajan *et al.*, 2010) (ArrayExpress: E-TABM-543), while the lower half (NKI) shows correlations for the dual-channel dataset of van de Vijver *et al.* (2002). Correlations were computed with the expression data as used by Weigelt *et al.* (2010a). A related but different view on the same data is offered in the supplementary web appendix of the reaction by Sørlie *et al.* (2010). The two BW plots in Panel (**A**) corresponding to the normal-like and luminal B centroid for the Natrajan dataset are completely non-overlapping. This implies that in this case the normal-like subtype will never be detected, since for each sample the correlation to the luminal B centroid is stronger. Furthermore, in the same panel one can observe that the range of the BW plots for each centroid is fairly small. In Panel (**B**), we see that when the data is properly scaled the BW plots are wider and more centered and hence they do not directly imply the exclusion of detection of a subtype. The lower halfs in panel (**A**) and (**B**) show the results for a dual-channel data set of log-ratio data from a common reference design. The impact of data centering is fairly small in this case. Comparing panel (**C**) to panel (**A**) we see that even when using uncentered data, the ranges obtained by PAM50 are wider than for the SSP of Sørlie, while for centered data (panel (**D**)) they are wider yet. In addition, the BW plots associated with PAM50 show much higher correlations compared to the SSP of Sørlie which suggests a larger confidence in subtype detection.

Panel	Index	Description	Type	κ (Weigelt)	κ (133plus2)	κ (all)	$\Delta\kappa$
All	1	weigelt uncentered	SSP.classic	0.467	-	-	-
	2	weigelt centered	SSP.classic	0.561	-	-	-
	3	SSP.classic.4s	SSP.classic	-	0.560	0.555	-0.005
	4	SCM.classic	SCM.classic	-	0.778	0.772	-0.006
	5	different CS and GL	SSP.cs	-	0.828	0.800	-0.028
	6	same CS, different GL	SSP.cs	-	0.854	0.820	-0.034
	7	same GL, different CS	SSP.cs	-	0.914	0.905	-0.009
	8	different CS and GL	SCM.cs	-	0.778	0.753	-0.025
	9	same CS, different GL	SCM.cs	-	0.814	0.776	-0.038
	10	same GL, different CS	SCM.cs	-	0.876	0.872	-0.004
Basal	1	weigelt uncentered	SSP.classic	0.921	-	-	-
	2	weigelt centered	SSP.classic	0.852	-	-	-
	3	SSP.classic.4s	SSP.classic	-	0.927	0.921	-0.006
	4	SCM.classic	SCM.classic	-	0.867	0.867	0.000
	5	different CS and GL	SSP.cs	-	0.948	0.937	-0.011
	6	same CS, different GL	SSP.cs	-	0.948	0.940	-0.008
	7	same GL, different CS	SSP.cs	-	0.987	0.979	-0.008
	8	different CS and GL	SCM.cs	-	0.906	0.892	-0.014
	9	same CS, different GL	SCM.cs	-	0.907	0.892	-0.015
	10	same GL, different CS	SCM.cs	-	0.987	0.982	-0.005
HER2	1	weigelt uncentered	SSP.classic	0.482	-	-	-
	2	weigelt centered	SSP.classic	0.564	-	-	-
	3	SSP.classic.4s	SSP.classic	-	0.502	0.480	-0.022
	4	SCM.classic	SCM.classic	-	0.762	0.795	+0.033
	5	different CS and GL	SSP.cs	-	0.808	0.798	-0.010
	6	same CS, different GL	SSP.cs	-	0.846	0.832	-0.014
	7	same GL, different CS	SSP.cs	-	0.916	0.901	-0.015
	8	different CS and GL	SCM.cs	-	0.756	0.754	-0.002
	9	same CS, different GL	SCM.cs	-	0.768	0.768	0.000
	10	same GL, different CS	SCM.cs	-	0.926	0.919	-0.007
LumA	1	weigelt uncentered	SSP.classic	0.449	-	-	-
	2	weigelt centered	SSP.classic	0.508	-	-	-
	3	SSP.classic.4s	SSP.classic	-	0.546	0.546	0.000
	4	SCM.classic	SCM.classic	-	0.766	0.757	-0.009
	5	different CS and GL	SSP.cs	-	0.808	0.776	-0.032
	6	same CS, different GL	SSP.cs	-	0.831	0.794	-0.037
	7	same GL, different CS	SSP.cs	-	0.904	0.897	-0.007
	8	different CS and GL	SCM.cs	-	0.750	0.732	-0.018
	9	same CS, different GL	SCM.cs	-	0.788	0.761	-0.027
	10	same GL, different CS	SCM.cs	-	0.840	0.839	-0.001
LumB	1	weigelt uncentered	SSP.classic	0.306	-	-	-
	2	weigelt centered	SSP.classic	0.244	-	-	-
	3	SSP.classic.4s	SSP.classic	-	0.276	0.311	+0.035
	4	SCM.classic	SCM.classic	-	0.701	0.678	-0.023
	5	different CS and GL	SSP.cs	-	0.738	0.715	-0.023
	6	same CS, different GL	SSP.cs	-	0.774	0.737	-0.037
	7	same GL, different CS	SSP.cs	-	0.857	0.857	0.000
	8	different CS and GL	SCM.cs	-	0.691	0.668	-0.023
	9	same CS, different GL	SCM.cs	-	0.736	0.702	-0.034
	10	same GL, different CS	SCM.cs	-	0.797	0.802	+0.005

	Weigelt uncentered	Weigelt centered	SSP.classic	SSP.classic.4s	SCM.classic	SSP.cs	SCM.cs
cc (all, %)	61.72	66.59	70.75 (71.26)	66.97 (67.01)	83.88 (83.66)	90.32 (88.99)	86.67 (85.33)
κ (all)	0.467	0.561	0.575 (0.570)	0.560 (0.555)	0.778 (0.772)	0.865 (0.842)	0.812 (0.791)
κ (basal)	0.921	0.852	0.914 (0.904)	0.927 (0.921)	0.867 (0.867)	0.970 (0.958)	0.930 (0.922)
κ (HER2)	0.482	0.564	0.500 (0.470)	0.502 (0.480)	0.762 (0.795)	0.846 (0.841)	0.802 (0.793)
κ (lumA)	0.449	0.508	0.642 (0.624)	0.546 (0.546)	0.766 (0.757)	0.845 (0.817)	0.778 (0.766)
κ (lumB)	0.306	0.244	0.276 (0.308)	0.276 (0.311)	0.701 (0.678)	0.780 (0.763)	0.731 (0.712)

Supplementary Table 3.3: Summary of intra-predictor concordance of SSPs and SCMs. Numerical details of Figure 3.3 in the main text: median percentage of concordant samples (cc) and median kappa statistics. Results presented in the first two columns were obtained using the published subtype assignments by Weigelt et al. on uncentered (Weigelt *et al.*, 2010a) and centered data (Weigelt *et al.*, 2010c), respectively. Remaining columns summarize results based on our hgu133plus2 compendium and - between parentheses - on the entire Affymetrix compendium, i.e. including the hgu133a arrays (Supplementary Table 3.1). For the CS-based predictors (SSP.cs, SCM.cs) the entries are pooled estimates corresponding to the vertical gray lines in Figure 3.3 in the main text (spanning 3 rows within each panel).

	Single SSP: centered vs. uncentered data				Pair of SSPs: centered data only				
	SSP	All	Single	Dual	SSP.1	SSP.2	All	Single	Dual
cc (all, %)	Sørlie	66.71	58.29	82.03	Sørlie	Hu	63.58	64.80	61.36
	Hu	68.87	63.50	78.64	Sørlie	PAM50	64.06	65.18	62.03
	PAM50	75.00	67.23	89.15	Hu	PAM50	72.36	72.25	72.54
κ (all)	Sørlie	0.570	0.462	0.766	Sørlie	Hu	0.528	0.541	0.504
	Hu	0.585	0.504	0.722	Sørlie	PAM50	0.532	0.544	0.510
	PAM50	0.677	0.575	0.861	Hu	PAM50	0.644	0.639	0.652
κ (basal)	Sørlie	0.879	0.840	0.962	Sørlie	Hu	0.792	0.813	0.744
	Hu	0.805	0.769	0.878	Sørlie	PAM50	0.894	0.899	0.882
	PAM50	0.903	0.902	0.905	Hu	PAM50	0.800	0.821	0.753
κ (HER2)	Sørlie	0.550	0.306	0.834	Sørlie	Hu	0.469	0.473	0.462
	Hu	0.650	0.564	0.820	Sørlie	PAM50	0.590	0.580	0.606
	PAM50	0.646	0.399	0.909	Hu	PAM50	0.584	0.541	0.650
κ (lumA)	Sørlie	0.648	0.573	0.779	Sørlie	Hu	0.518	0.520	0.516
	Hu	0.516	0.454	0.638	Sørlie	PAM50	0.443	0.458	0.417
	PAM50	0.664	0.560	0.854	Hu	PAM50	0.629	0.627	0.634
κ (lumB)	Sørlie	0.330	0.266	0.499	Sørlie	Hu	0.197	0.200	0.194
	Hu	0.538	0.376	0.741	Sørlie	PAM50	0.217	0.197	0.252
	PAM50	0.554	0.432	0.852	Hu	PAM50	0.624	0.611	0.645
κ (normal)	Sørlie	0.412	0.073	0.793	Sørlie	Hu	0.581	0.592	0.564
	Hu	0.289	0.028	0.543	Sørlie	PAM50	0.493	0.532	0.430
	PAM50	0.630	0.529	0.749	Hu	PAM50	0.500	0.457	0.557
Nr. of samples	-	832	537	295	-	-	832	537	295

Supplementary Table 3.4: Concordance comparison: centering vs. difference in classic SSP. Percentage of concordant samples (cc) and kappa statistics calculated using subtype assignments by Weigelt *et al.* (2010a,c) (`http://rock.icr.ac.uk/collaborations/Mackay/centroid.correlations.Eset/`). The left-hand half of the table lists the concordance for a single classic SSP between assignments based on uncentered data and on centered data. In the right-hand half all comparisons were based on centered data for pairs of different classic SSPs. The columns *All*, *Single* and *Dual* indicate results that were computed on all samples, only samples from single-channel experiments or only samples from dual-channel experiments, respectively.

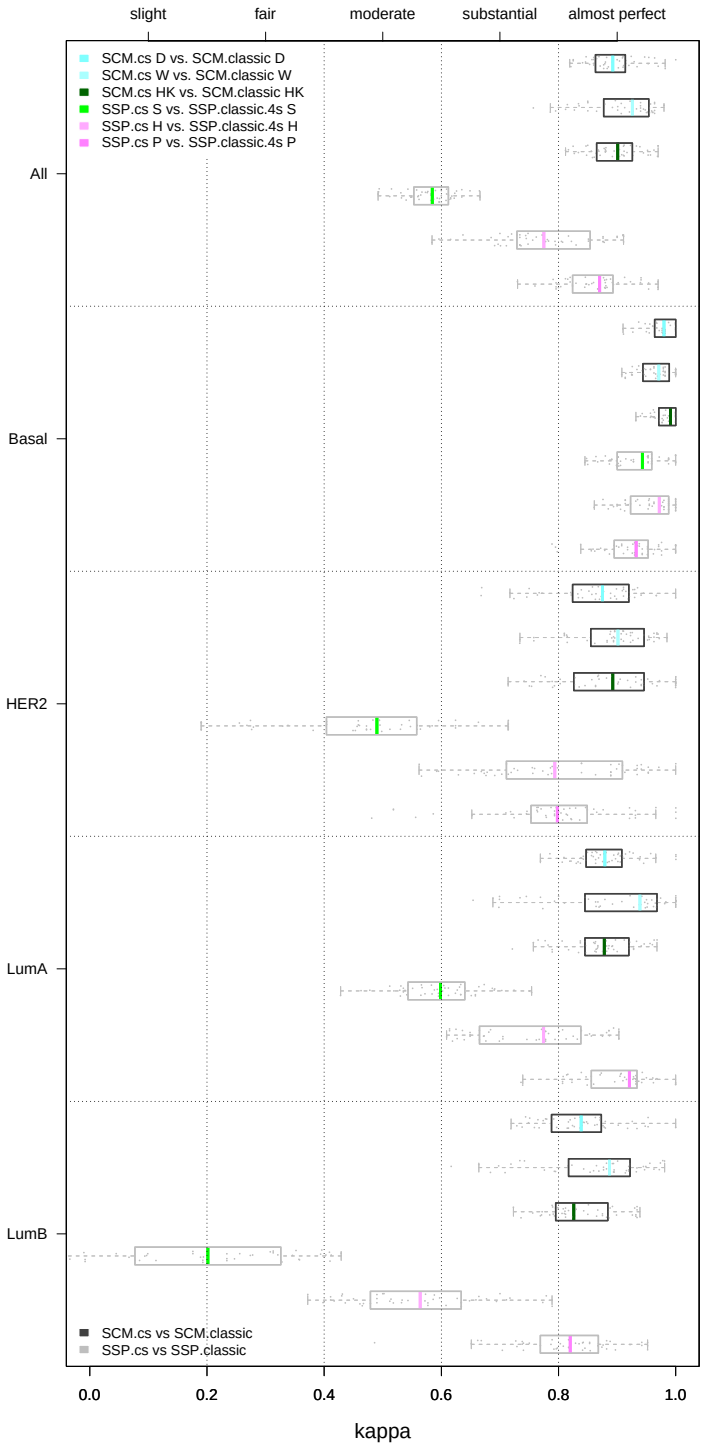
	Weigelt (centered)	Haibe-Kains	Guedj	Perou
cc (%, Sørlie vs. Hu)	63.58	69.25	58.24	
cc (%, Sørlie vs. PAM50)	64.06	60.13	57.64	
cc (%, Hu vs. PAM50)	72.36	73.64	75.74	77.60
κ (all, Sørlie vs. Hu)	0.528	0.581	0.464	
κ (all, Sørlie vs. PAM50)	0.532	0.474	0.460	
κ (all, Hu vs. PAM50)	0.644	0.652	0.690	0.710
κ (basal, Sørlie vs. Hu)	0.792	0.906	0.692	
κ (basal, Sørlie vs. PAM50)	0.894	0.907	0.750	
κ (basal, Hu vs. PAM50)	0.800	0.906	0.771	0.856
κ (HER2, Sørlie vs. Hu)	0.469	0.430	0.347	
κ (HER2, Sørlie vs. PAM50)	0.590	0.556	0.420	
κ (HER2, Hu vs. PAM50)	0.584	0.700	0.648	0.722
κ (lumA, Sørlie vs. Hu)	0.518	0.617	0.456	
κ (lumA, Sørlie vs. PAM50)	0.443	0.372	0.396	
κ (lumA, Hu vs. PAM50)	0.629	0.606	0.704	0.673
κ (lumB, Sørlie vs. Hu)	0.197	0.192	0.275	
κ (lumB, Sørlie vs. PAM50)	0.217	0.196	0.257	
κ (lumB, Hu vs. PAM50)	0.624	0.476	0.633	0.618
κ (normal, Sørlie vs. Hu)	0.581	0.541	0.540	
κ (normal, Sørlie vs. PAM50)	0.493	0.314	0.530	
κ (normal, Hu vs. PAM50)	0.500	0.451	0.682	0.588
Nr. of samples	832	2576	2828	442

Supplementary Table 3.5: Lack of concordance of classic SSPs on large sample sizes. Percentage of concordant samples (cc) and kappa statistics for pairs of classic SSPs, based on recently published subtype assignments from four research groups. The top row indicates the different sources from which subtype assignments were obtained. *Weigelt (centered)*: subtype assignments from the rebuttal by Weigelt *et al.* (2010c); *Haibe-Kains*: subtype assignments by Haibe-Kains *et al.* (2012); *Guedj*: subtype assignments by Guedj *et al.* (2011); *Perou*: subtype assignments from the Perou group, as listed on the UNC website (`https://genome.unc.edu/pubsup/breastGEO/`). All subtype assignments were based on scaled expression data. Most experiments were based on median-centered data, however, Haibe-Kains *et al.* (2012) relied on robust scaling (Materials and Methods, main text). Subtype assignments by the Perou group were only available for the SSP of Hu and PAM50.

Platform	All	Others	Affy	All	Others	Affy	All	Others	Affy
SCM.1	HK	HK	HK	HK	HK	HK	D	D	D
SCM.2	D	D	D	W	W	W	W	W	W
cc (all, %)	74.83	65.72	**81.91**	78.01	69.45	**84.67**	85.66	82.40	**88.24**
κ (all)	0.662	0.542	**0.756**	0.703	0.590	**0.792**	0.807	0.763	**0.841**
κ (basal)	0.715	0.546	**0.860**	0.749	0.574	**0.898**	0.897	0.854	**0.930**
κ (HER2)	0.730	0.664	**0.787**	0.811	0.735	**0.877**	0.766	0.726	**0.800**
κ (lumA)	0.646	0.538	**0.716**	0.675	0.578	**0.741**	0.804	0.765	**0.831**
κ (lumB)	0.576	0.448	**0.677**	0.609	0.510	**0.685**	0.754	0.706	**0.794**
Nr. of samples	4606	2030	2576	4606	2030	2576	4606	2030	2576

Supplementary Table 3.6: Impact of platform heterogeneity on SCM-based subtyping. Percentage of concordant samples (cc) and kappa statistics for pairs of classic SCMs. Individual subtype assignments as published in Haibe-Kains *et al.* (2012) were aggregated into a single vector and subsequently dichotomized into a set containing only assignments made on Affymetrix datasets and another set made on non-Affymetrix datasets. *SCM.1* and *SCM.2* indicate for each pair the classic SCMs involved (HK, D, W); *Platform* indicates the set of experiments on which levels of concordance were computed; *All* refers to all subtype assignments, irrespective of the corresponding measurement platform; *Affy* refers to subtype assignments based on data from the Affymetrix platform only; *Others* refers to subtype assignments for all non-Affymetrix platforms; *Nr. of samples* indicates the total number of arrays in each sub-analysis. Bold entries indicate the best performance for a given SCM pair.

Supplementary Figure 3.2 *(facing page)*: Concordance of CS-based models and their classic counterparts (hgu133plus2 compendium). The five panels show box and whisker (BW) plots for kappa statistics calculated over all subtypes and for each subtype separately, as indicated on the left hand side. Results for individual datasets are superimposed as dots. Each panel contains six BW plots. The upper three BW plots indicate concordance for SCM predictor pairs, whereas the lower three BW plots indicate concordance for SSP predictor pairs. For the classic predictors we used the *genefu* implementations, i.e. *ssp2003.robust* (SSP.classic.4s.S), *ssp2006.robust* (SSP.classic.4s.H), *pam50.robust* (SSP.classic.4s.P), *scmod1.robust* (SCM.classic.D), *scmod2.robust* (SCM.classic.W) and *scmgene.robust* (SCM.classic.HK). For each predictor pair, the CS version used the same gene list (IGL for SSPs, MGL for SCMs) as its classic counterpart. Results are based on the hgu133plus2 compendium. Top legend: composition of each pair indicated by the color of BW median values (indicated by a bar). Bottom legend: predictor type indicated by the color of a BW box. Numerical details of the BW plots are presented in Supplementary Table 3.7.


slight
fair
moderate
substantial
almost perfect
SCM.cs D vs. SCM.classic D
SCM.cs W vs. SCM.classic W
SCM.cs HK vs. SCM.classic HK
SSP.cs S vs. SSP.classic.4s S
SSP.cs H vs. SSP.classic.4s H
SSP.cs P vs. SSP.classic.4s P
All
Basal
HER2
LumA
LumB
SCM.cs vs SCM.classic
SSP.cs vs SSP.classic
0.0
0.2
0.4
0.6
0.8
1.0
kappa

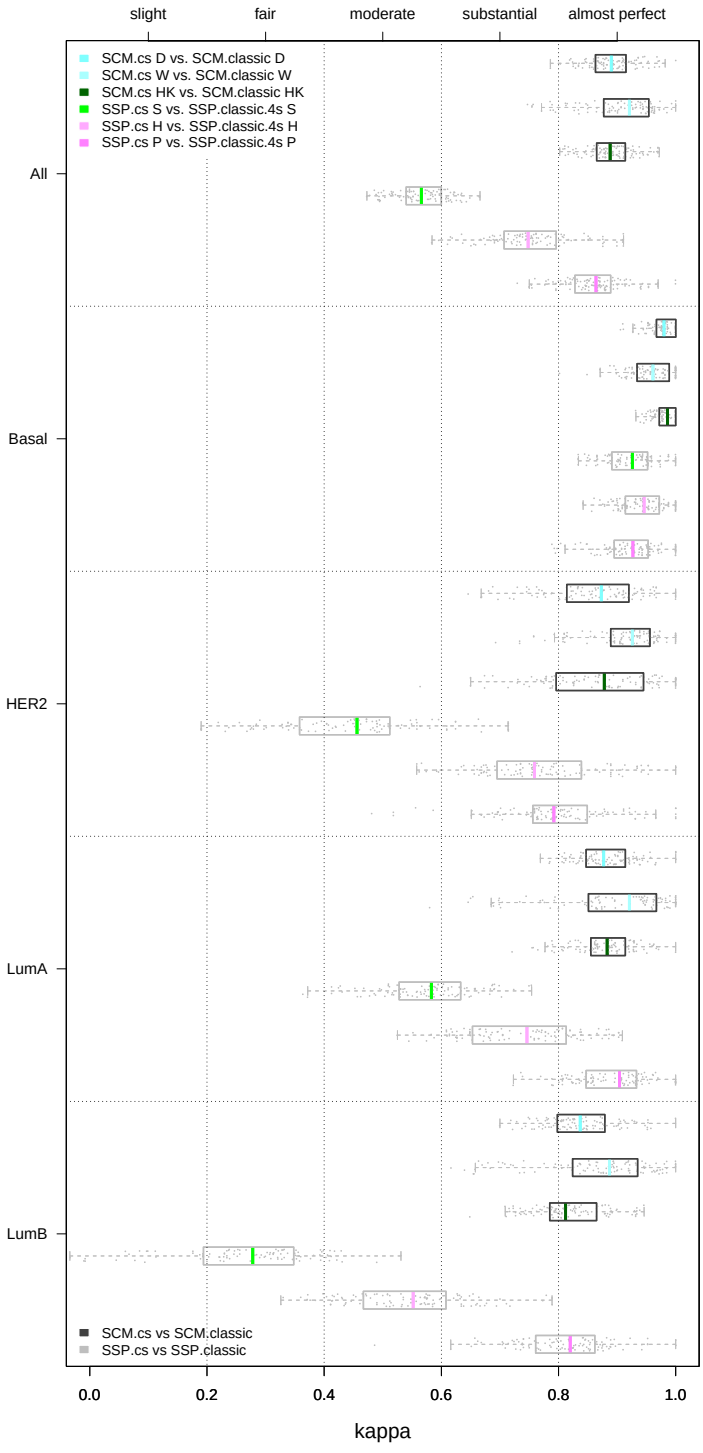
	Description	hgu133plus2	All	Δ
cc (all, %)	SCM.cs D vs. SCM.classic D	92.15	92.13	-0.02
	SCM.cs W vs. SCM.classic W	94.55	94.42	-0.13
	SCM.cs HK vs. SCM.classic HK	92.89	92.42	-0.47
	SSP.cs S vs. SSP.classic S	70.24	70.08	-0.16
	SSP.cs H vs. SSP.classic H	83.95	82.22	-1.73
	SSP.cs P vs. SSP.classic P	90.77	90.20	-0.57
κ (all)	SCM.cs D vs. SCM.classic D	0.893	0.890	-0.003
	SCM.cs W vs. SCM.classic W	0.926	0.921	-0.005
	SCM.cs HK vs. SCM.classic HK	0.901	0.888	-0.013
	SSP.cs S vs. SSP.classic S	0.584	0.566	-0.018
	SSP.cs H vs. SSP.classic H	0.775	0.748	-0.027
	SSP.cs P vs. SSP.classic P	0.870	0.864	-0.006
κ (basal)	SCM.cs D vs. SCM.classic D	0.980	0.980	0
	SCM.cs W vs. SCM.classic W	0.971	0.961	-0.01
	SCM.cs HK vs. SCM.classic HK	0.991	0.986	-0.005
	SSP.cs S vs. SSP.classic S	0.943	0.926	-0.017
	SSP.cs H vs. SSP.classic H	0.972	0.946	-0.026
	SSP.cs P vs. SSP.classic P	0.933	0.927	-0.006
κ (HER2)	SCM.cs D vs. SCM.classic D	0.875	0.873	-0.002
	SCM.cs W vs. SCM.classic W	0.902	0.926	+0.024
	SCM.cs HK vs. SCM.classic HK	0.893	0.878	-0.015
	SSP.cs S vs. SSP.classic S	0.490	0.456	-0.034
	SSP.cs H vs. SSP.classic H	0.794	0.759	-0.035
	SSP.cs P vs. SSP.classic P	0.798	0.792	-0.006
κ (lumA)	SCM.cs D vs. SCM.classic.D	0.879	0.877	-0.002
	SCM.cs W vs. SCM.classic W	0.939	0.921	-0.018
	SCM.cs HK vs. SCM.classic HK	0.878	0.883	+0.005
	SSP.cs S vs. SSP.classic S	0.598	0.583	-0.015
	SSP.cs H vs. SSP.classic H	0.774	0.746	-0.028
	SSP.cs P vs. SSP.classic P	0.921	0.904	-0.017
κ (lumB)	SCM.cs D vs. SCM.classic D	0.838	0.837	-0.001
	SCM.cs W vs. SCM.classic W	0.887	0.887	0
	SCM.cs HK vs. SCM.classic HK	0.826	0.812	-0.014
	SSP.cs S vs. SSP.classic S	0.202	0.278	+0.076
	SSP.cs H vs. SSP.classic H	0.564	0.552	-0.012
	SSP.cs P vs. SSP.classic P	0.820	0.820	0

Supplementary Table 3.7: Concordance of CS-based models and their classic counterparts (hgu133plus2 and hgu133a compendium). The upper six rows provide the concordance percentages between CS-based SSPs/SCMs and their classic counterparts. The remaining rows correspond to the median kappa statistics shown in Supplementary Figure 3.2 (column *hgu133plus2*) and Supplementary Figure 3.3 (column *All*). Δ: difference between median concordance percentages or median kappa statistics computed on hgu133plus2 samples and on the complete compendium, respectively.

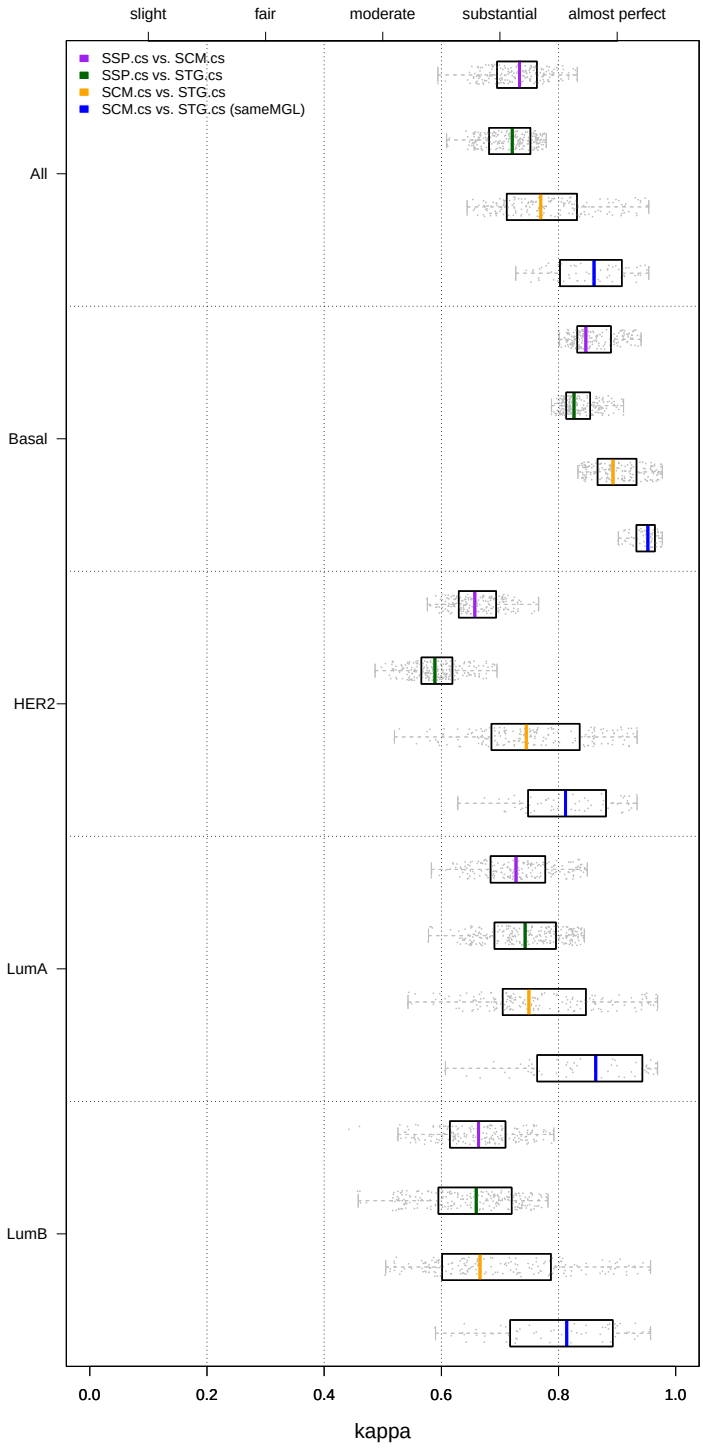
Description	cc (all, %)	κ (all)	κ (basal)	κ (HER2)	κ (lumA)	κ (lumB)
SSP.cs vs. SCM.cs	81.02 (80.69)	0.741 (0.734)	0.849 (0.847)	0.671 (0.657)	0.734 (0.728)	0.688 (0.663)
SSP.cs vs. STG.cs	80.07 (79.92)	0.729 (0.721)	0.826 (0.827)	0.599 (0.589)	0.755 (0.743)	0.682 (0.659)
SCM.cs vs. STG.cs	83.79 (83.32)	0.780 (0.770)	0.893 (0.893)	0.742 (0.745)	0.752 (0.750)	0.718 (0.666)
SCM.cs vs. STG.cs (same MGL)	89.84 (89.95)	0.861 (0.861)	0.953 (0.952)	0.811 (0.812)	0.853 (0.863)	0.811 (0.814)

Supplementary Table 3.8: Inter-predictor concordance of CS-based models (hgu133plus2 and hgu133a compendium). Numerical details of Figure 3.4 in the main text: median percentage of concordant samples (cc) and median kappa statisticss between the SSP.cs, SCM.cs and STG.cs predictor pairs as indicated in the first column. Entries between parentheses represent results over the entire Affymetrix compendium, i.e. including the hgu133a arrays (Supplementary Figure 3.4).

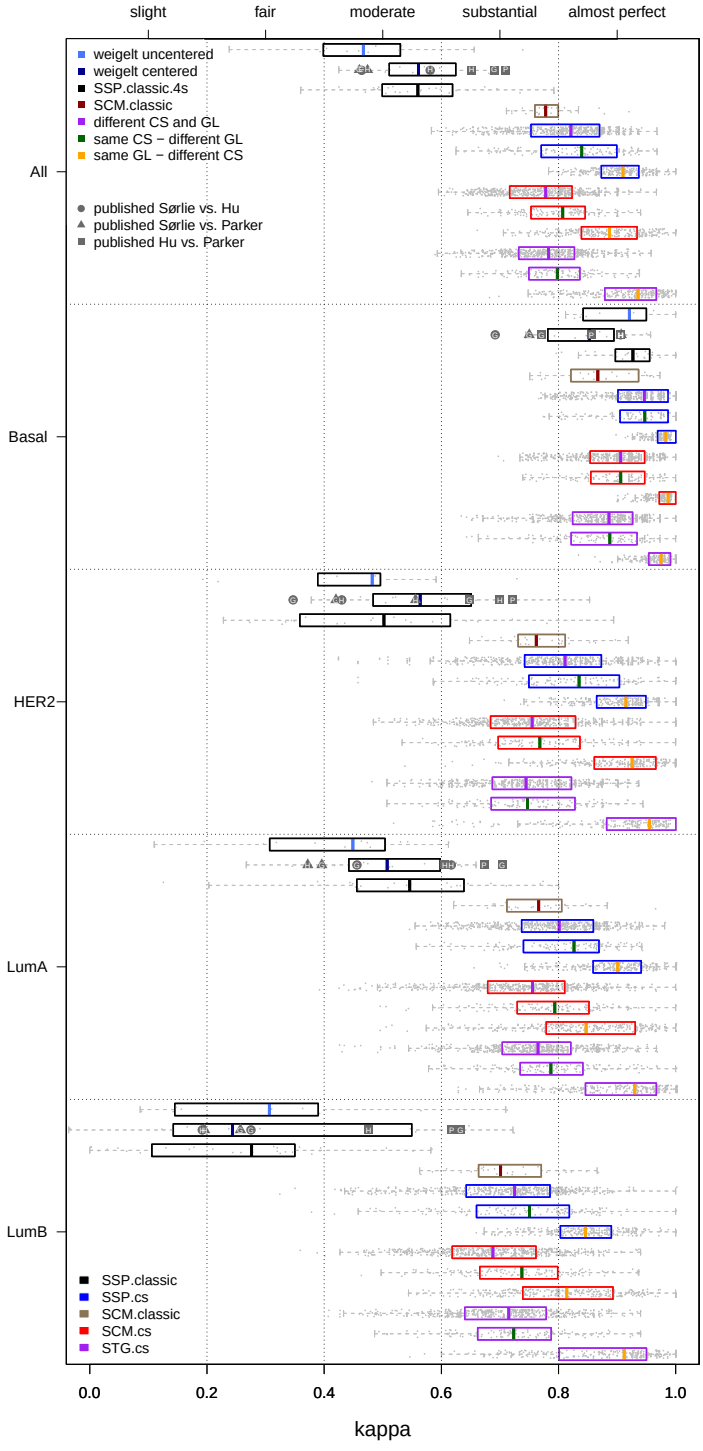
Supplementary Figure 3.3 *(facing page)*: Concordance of CS-based models and their classic counterparts (hgu133plus2 and hgu133a compendium). Complement to Supplementary Figure 3.2. The five panels show box and whisker (BW) plots for kappa statistics calculated over all subtypes and for each subtype separately, as indicated on the left hand side. Results for individual datasets are superimposed as dots. Each panel contains six BW plots. The upper three BW plots indicate concordance for SCM predictor pairs, whereas the lower three BW plots indicate concordance for SSP predictor pairs. For each predictor pair, the CS version used the same gene list (IGL for SSPs, MGL for SCMs) as its classic counterpart. Results are based on the entire Affymetrix compendium. Top legend: composition of each pair indicated by the color of BW median values (indicated by a bar). Bottom legend: predictor type indicated by the color of a BW box. Numerical details of the BW plots are presented in Supplementary Table 3.7.

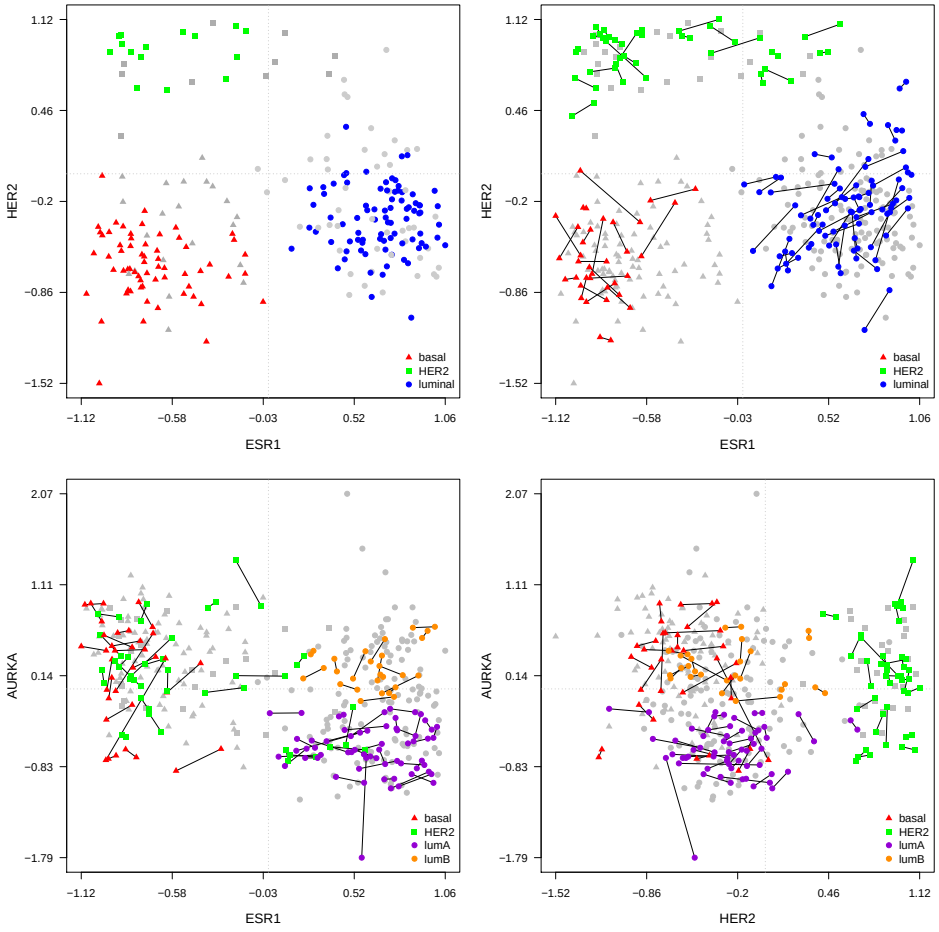

slight
fair
moderate
substantial
almost perfect
SCM.cs D vs. SCM.classic D
SCM.cs W vs. SCM.classic W
SCM.cs HK vs. SCM.classic HK
SSP.cs S vs. SSP.classic.4s S
SSP.cs H vs. SSP.classic.4s H
SSP.cs P vs. SSP.classic.4s P
All
Basal
HER2
LumA
LumB
SCM.cs vs SCM.classic
SSP.cs vs SSP.classic
0.0
0.2
0.4
0.6
0.8
1.0
kappa

Supplementary Figure 3.4 ***(facing page)*****: Inter-predictor concordance of CS-based models (hgu133plus2 and hgu133a compendium)**. Complement to Figure 3.4 in the main text. The five panels show box and whisker plots for kappa statistics calculated over all subtypes and for each each subtype separately, as indicated on the left hand side. Results for individual datasets are superimposed as dots. The upper three BW plots in each panel show the inter-predictor concordance estimates between the SSP.cs, SCM.cs and STG.cs predictors pairs, as indicated by the legend. The bottom BW plot in each panel provides the concordance estimates for SCM.cs and STG.cs predictor pairs when based on the same modules, i.e. MGLs (with exception of PGR). Results are based on the entire Affymetrix compendium. Numerical details of the BW plots are presented in Supplementary Table 3.8.


slight
fair
moderate
substantial
almost perfect
SSP.cs vs. SCM.cs
SSP.cs vs. STG.cs
SCM.cs vs. STG.cs
SCM.cs vs. STG.cs (sameMGL)
All
Basal
HER2
LumA
LumB
0.0
0.2
0.4
0.6
0.8
1.0
kappa

Supplementary Figure 3.5 ***(facing page)*****: Intra-predictor concordance of SSPs, SCMs and STGs (hgu133plus2 compendium)**. Complement to Figure 3.3 in the main text. Also provides intra-STG.cs concordance estimates in addition to SSP and SCM intra-predictor concordance estimates. The modules used in the STG.cs model were identical to those used for the SCMs (with exception of PGR). Results were based on our hgu133plus2 compendium. See the legend of Figure 3.3 for more detailed information.


slight
fair
moderate
substantial
almost perfect
weigelt uncentered
weigelt centered
SSP.classic.4s
SCM.classic
different CS and GL
same CS − different GL
same GL − different CS
published Sørlie vs. Hu
published Sørlie vs. Parker
published Hu vs. Parker
All
Basal
HER2
LumA
LumB
SSP.classic
SSP.cs
SCM.classic
SCM.cs
STG.cs
0.0
0.2
0.4
0.6
0.8
1.0
kappa



Supplementary Figure 3.6: Concordance on replicate array pairs. SCM.cs HK was fitted on the 162 consensus set samples (Table 3.1, main text) from the Sabatier cohort (Dataset D9, Supplementary Table 3.1). The top left panel shows the module score scatterplot for the ESR1 and HER2 modules for all 266 samples from the Sabatier cohort. Consensus set samples are highlighted in color, whereas the remaining samples are shown in gray. The other panels show the module scores for a set of 93 replicate array pairs in our compendium. In these the module scores from the top left panel are shown in gray, while the module scores for the replicate arrays are highlighted in color. Module scores of replicate array pairs are connected by a line segment. In total 86 out of 93 replicate pairs were consistently subtyped by this predictor (κ=0.90, cc=92.47%). For clarity only consistently subtyped pairs are shown.

Subset	cc (all, %)	κ (all)	κ (basal)	κ (HER2)	κ (lumA)	κ (lumB)
All	91.40	0.883	0.894	0.968	0.847	0.844
SSP.cs	91.40	0.883	0.881	0.969	0.859	0.868
SCM.cs	91.40	0.884	0.940	0.968	0.844	0.856

Supplementary Table 3.9: Concordance on replicate array pairs. Percentage of concordant samples (cc) and kappa statistics for a set of 93 replicate array pairs in our compendium. Concordance values are median values over all pairs of CS-based predictors (*All*), SSP.cs predictors and SCM.cs predictors, respectively.

	ER HK	ER D	ER W	HER2 HK	HER2 D	HER2 W	PGR	Proliferation	AURKA HK	AURKA D	AURKA W
Bos	2.44	1.97	1.86	1.80	1.07	2.03	1.69	0.75	0.61	1.05	1.02
Dedeurwaerder	2.43	1.97	2.09	2.65	1.29	2.23	2.15	1.24	1.03	1.14	1.12
Desmedt	2.62	1.90	1.85	1.31	0.79	1.61	1.37	1.22	1.20	1.30	1.38
expO	2.39	1.37	1.53	1.46	1.12	1.59	1.57	1.36	1.45	1.36	1.29
Farmer	2.95	2.13	1.77	3.41	1.37	2.27	1.92	1.17	1.32	1.17	1.07
Guedj	2.12	1.49	1.49	1.45	0.89	1.65	1.73	0.91	1.14	1.01	0.91
Kao	1.88	1.33	1.47	1.76	1.07	1.70	1.43	1.08	1.16	1.16	1.10
Li	3.12	2.01	2.02	1.16	1.05	1.56	1.52	1.50	1.35	1.21	1.24
Lu	3.26	2.16	2.14	1.24	1.02	1.49	1.35	1.53	1.41	1.41	1.44
Miller	2.10	1.25	1.26	1.78	1.12	1.83	1.44	1.42	1.19	1.45	1.39
MSK	2.75	1.88	2.03	1.30	1.24	1.71	1.29	0.95	1.43	1.24	1.03
Pawitan	1.88	1.41	1.46	1.18	1.08	1.34	1.11	1.46	1.40	1.36	1.28
Richardson (I)	2.24	1.69	1.43	1.64	1.13	1.99	1.93	1.55	1.38	1.45	1.50
Richardson (II)	2.97	2.25	2.24	1.67	1.35	1.62	2.66	1.55	1.10	1.15	1.26
Sabatier	2.45	1.93	2.09	1.28	0.98	1.59	1.83	1.30	1.17	1.14	1.14
Schmidt	2.03	1.46	1.40	1.32	1.19	1.67	1.56	0.97	1.27	1.00	1.02
Shi	2.42	1.52	1.56	1.51	1.25	1.69	1.63	1.27	0.88	1.20	1.19
Symmans (I,II) + VDX	2.18	1.55	1.52	1.42	1.09	1.68	1.49	1.15	1.37	1.04	1.10
Symmans (III) + expO	2.34	1.34	1.47	1.34	0.98	1.54	1.47	1.35	1.47	1.29	1.23
UNT	2.05	1.58	1.57	1.63	1.06	1.69	1.36	1.36	1.23	1.56	1.43
VDX	2.55	1.83	1.80	1.61	1.36	1.90	1.49	1.23	1.29	1.21	1.21
BMI (median)	2.42	1.69	1.57	1.46	1.09	1.68	1.52	1.27	1.27	1.21	1.21
Nr. BMI $\geq$ 1.1	21	21	21	21	10	21	21	16	18	17	15
Nr. BMI $\geq$ 1.5	21	14	14	10	0	19	11	4	0	1	0

Supplementary Table 3.10: Bimodality indices of individual modules (hgu133plus2 and hgu133a compendium). Wang *et al.* (2009) characterize a distribution as being bimodal if the bimodality index (BMI) $\geq$ 1.1 and strongly bimodal if BMI $\geq$ 1.5. The first row indicates the various modules used to measure ER, HER2, PGR and proliferation (Supplementary Information). Proliferation was measured by the AURKA proliferation modules by Haibe-Kains *et al.* (2012) (HK), Desmedt *et al.* (2008) (D) and Wirapati *et al.* (2008) (W) and the proliferation module (Proliferation) (see Supplementary Information). BMI values are listed for each dataset from the Affymetrix compendium. The last three rows provide the median BMI value over all 21 datasets, the number of times the module was bimodal and the number of times the module was strongly bimodal, respectively.

Profile	ER	HER2	KI-67	PGR	5 Subtype STG	4 Subtype STG
1	ER-	HER2-	KI-67-	PGR-	basal	basal
2	ER-	HER2-	KI-67-	PGR+	luminal A	luminal A
3	ER-	HER2-	KI-67+	PGR-	basal	basal
4	ER-	HER2-	KI-67+	PGR+	luminal B (HER2-)	luminal B
5	ER-	HER2+	KI-67-	PGR-	HER2	HER2
6	ER-	HER2+	KI-67-	PGR+	luminal B (HER2+)	HER2
7	ER-	HER2+	KI-67+	PGR-	HER2	HER2
8	ER-	HER2+	KI-67+	PGR+	luminal B (HER2+)	HER2
9	ER+	HER2-	KI-67-	PGR-	luminal A	luminal A
10	ER+	HER2-	KI-67-	PGR+	luminal A	luminal A
11	ER+	HER2-	KI-67+	PGR-	luminal B (HER2-)	luminal B
12	ER+	HER2-	KI-67+	PGR+	luminal B (HER2-)	luminal B
13	ER+	HER2+	KI-67-	PGR-	luminal B (HER2+)	HER2
14	ER+	HER2+	KI-67-	PGR+	luminal B (HER2+)	HER2
15	ER+	HER2+	KI-67+	PGR-	luminal B (HER2+)	HER2
16	ER+	HER2+	KI-67+	PGR+	luminal B (HER2+)	HER2

A

Profile	ER	HER2	AURKA	SCM
i	ER-	HER2-	AURKA-	basal
ii	ER-	HER2-	AURKA+	basal
iii	ER-	HER2+	AURKA-	HER2
iv	ER-	HER2+	AURKA+	HER2
v	ER+	HER2-	AURKA-	luminal A
vi	ER+	HER2-	AURKA+	luminal B
vii	ER+	HER2+	AURKA-	HER2
viii	ER+	HER2+	AURKA+	HER2

B

Supplementary Figure 3.7: Molecular taxonomy of the St. Gallen and SCM subtyping schemes. **A**) The St. Gallen subtype defitions (Goldhirsch *et al.*, 2011) present a subtyping scheme based on the over(+)/under(-)expression of clinical markers for ER, HER2, KI-67 (prolifation status) and PGR. These four markers allow for $2^4 = 16$ distinct profiles. Each row corresponds to a particular profile, columns *ER*, *HER2*, *KI-67* and *PGR* indicate the over/underexpression status for each marker. For clarity, differences between groups are also highlighted by color. The column *5 Subtype STG* indicates the St. Gallen surrogate definitions of the intrinsic subtypes for each profile. In these, the luminal B subtype is subdivided into a luminal B (HER+) and luminal B (HER2-) subtype, while the HER2 subtype is associated with ER- profiles only. The column *4 Subtype STG* presents a mapping to the four main subtypes considered in this chapter. Note that in this case we made the deliberate decision to map the luminal B (HER2-) profiles to the luminal B subtype, while the luminal B (HER2+) profiles were mapped to the HER2 subtype. This mapping was chosen in order to maximize similarity with SCMs as shown in Panel (**B**). SCMs do not consider PGR status and therefore lead to $2^3 = 8$ distinct profiles. For comparison the profiles (rows) are ordered in the same way as the STG profiles. Assuming all processes are measured in the same way, in most cases, input vectors with identical ER, HER2 and proliferation status will be mapped to the same subtype. Discordance, however, may arise due to PGR status. For a sample with an identical ER, HER2 and proliferation profile, a luminal A or B subtype may be obtained for STGs (panel (**A**), profiles (2) and (4)), while for SCMs a basal subtype is obtained (panel (**B**), profiles (1) and (2)). Furthermore, the actual level of over/underexpression of a marker is relevant for SCMs, but not for STGs. This is likely to introduce additional discordance.

3.7 Supplementary Information

3.7.1 Subtype predictors

This section provides a comprehensive description and references to the literature for the classes of different subtype predictors used in the main manuscript.

SSP: single sample predictor

The classic single sample predictors are nearest centroid predictors, that is, prototype-driven classification rules (Friedman *et al.*, 2001) that are completely defined by a set of centroids and a suitable distance function (Figure 3.1A, main text). In line with previously described SSP schemes (Hu *et al.*, 2006; Parker *et al.*, 2009), we used the Spearman rank correlation distance measure. SSPs were constructed using the intrinsic gene lists (IGLs) related to the classic SSPs. We refer to the IGLs of the SSPs by Sørlie *et al.* (2003), Hu *et al.* (2006) and Parker *et al.* (2009) as the IGL S, H and P, respectively. For the classic SSPs we used the following functions from the *genefu* package: *ssp2003.robust* (SSP Sørlie), *ssp2006.robust* (SSP Hu) and *pam50.robust* (SSP Parker).

SCM: subtype classification model

As an alternative to SSPs, Desmedt *et al.* (2008) proposed a biology-inspired module-driven approach referred to as subtype classification models (Haibe-Kains *et al.*, 2012) (Figure 3.1B, main text). Module scores are calculated for three modules that reflect the activity of several key biological processes: (*i*) estrogen receptor signaling, (*ii*) HER2 signaling and (*iii*) proliferation. Three SCMs have been published previously, based on the same set of prototypes: the SCM by Desmedt *et al.* (2008), the SCM by Wirapati *et al.* (2008) and more recently the SCM by Haibe-Kains *et al.* (2012), also known as SCMGENE. We refer to these as the classic SCMs. In addition, for a given SCM we refer to the list of genes associated with a module as the module gene list (MGL). The latter can be thought of as the SCM equivalent of an IGL. We refer to the MGLs corresponding to the SCMs by Desmedt *et al.* (2008), Wirapati *et al.* (2008) and Haibe-Kains *et al.* (2012) as the MGLs D, W and HK, respectively. For the classic SCMs we used the following functions from the *genefu* package: *scmod1.robust* (SCM Desmedt), *scmod2.robust* (SCM Wirapati) and *scmgene.robust* (SCM Haibe-Kains).

Probeset	HUGO gene symbol	Entrez Gene ID
202095_s_at	BIRC5	332
202589_at	TYMS	7298
202870_s_at	CDC20	991
202954_at	UBE2C	11065
209773_s_at	RRM2	6241
214710_s_at	CCNB1	891

Supplementary Table 3.11: STG proliferation module. The module composition of the 6-gene proliferation module was based on the intersection of all genes in the AURKA proliferation modules by Desmedt *et al.* (2008) and Wirapati *et al.* (2008) retrieved from the *genefu* package and the 11-gene proliferation signature proposed by Nielsen *et al.* (2010). The latter signature consists of the HUGO gene symbol entries: CCNB1, UBE2C, BIRC5, KNTC2, CDC20, PTTG1, RRM2, MKI67, TYMS, CEP55, CDCA1. All probesets had a weight of +1 in the calculation of the module score.

STG: predictor based on St. Gallen surrogate intrinsic subtypes

In this study, we developed a rule-based predictor (STG) derived from the St. Gallen surrogate intrinsic subtype definitions which are based on clinical markers of ER, HER2, PGR and KI-67 (proliferation) status (Goldhirsch *et al.*, 2011). An STG is fully defined by the over/underexpression status of the markers, which allows for 16 distinct profiles (Figure 3.1C, main text). Over/underexpression status of the four markers was determined by considering module scores. The ER, HER2 and PGR modules consisted of a single probeset. These correspond to the probesets previously suggested for these processes (Wang *et al.*, 2009), and for ER and HER2 are identical to those used by SCMGENE. The proliferation module was based on the intersection of all genes in the AURKA proliferation modules by Desmedt and Wirapati and the 11-gene proliferation signature proposed by Nielsen *et al.* (2010). This resulted in a 6-gene proliferation module (Supplementary Table 3.11). For each marker and training set separately, over/underexpression was estimated by fitting a 2-component Gaussian mixture model on the module scores. For each component i, let u_i, σ_i^2 and w_i be the estimated mean, variance and mixing proportion, respectively. Assuming equal variances, the following cutoff can be used to determine the actual over/underexpression status for a new case:

$$c = \frac{\sigma^2 log(w_2/w_1) + 2(u_1^2 - u_2^2)}{u_1 - u_2}.$$

Cases with a module score larger than or equal to c were considered overexpressed, while the others were considered underexpressed.

3.7.2 Consensus sets

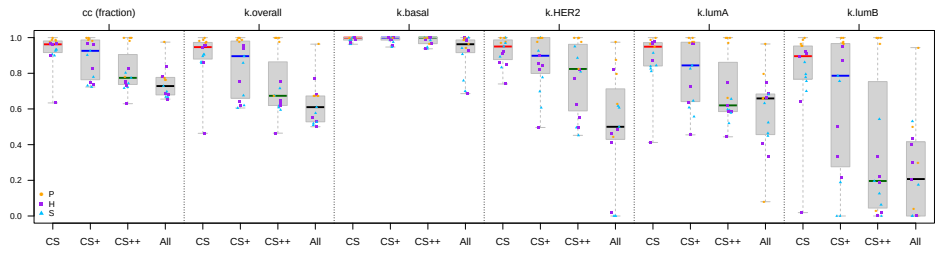
This section gives an overview of a number of additional experiments, characterizing the consensus set samples in more detail.

Consensus set subtype identification by hierarchical clustering

In breast cancer literature SSP construction has almost always been linked to unsupervised learning via hierarchical clustering (HC) (Sørlie *et al.*, 2003; Hu *et al.*, 2006; Parker *et al.*, 2009; Guedj *et al.*, 2011). Instability of hierarchical clustering is a well-known problem (Tibshirani and Walther, 2005; Lusa *et al.*, 2007). Haibe-Kains *et al.* (2012) reported very low levels of concordance for HC-based SSP predictors when clustering complete sample cohorts. We investigated to what extent the subtype labels of the consensus sets could have been identified by HC alone and to what degree their identification is influenced by the presence of additional samples during clustering. Importantly, for any given dataset concordance was always measured over the CS samples only. When we only cluster CS samples, in all but one case almost perfect levels of concordance were obtained (Supplementary Figure 3.8). However, it becomes increasingly more difficult to identify the CS subtype labels by HC when the training set becomes larger (and more heterogeneous). Furthermore, similar to Pusztai *et al.* (2006), results strongly depended on the selected IGL. For the IGL P in nearly all cases almost perfect levels of concordance were obtained, however, not when clustering the CS samples in the presence of all additional samples. Concordance for the IGLs H and S was notably lower, especially when clustering CS samples in the presence of additional samples. Lowest concordance was observed for the luminal B subtype, whose concordance with CS subtype labels decreased strongly in the presence of additional samples.

Bimodality status of individual modules

Module scores are a core ingredient of both SCMs and STGs (Section 3.7.1). For a module score that is unimodally distributed, it is difficult to estimate a sensible cutoff for determining the over/underexpression status of the module for individual cases. The bimodality status of a module score, therefore, provides a good indication of the performance of SCM and STG subtyping schemes. We used the bimodality index (BMI) (Wang *et al.*, 2009) to assess bimodality of the distribution of the module scores related to ER, HER2, and PGR signaling and proliferation on the five consensus sets (Supplementary Table 3.13). In most instances all modules showed strong indications of bimodality (BMI$\geq$1.5). However, the level of bimodality depended on both the dataset and module composition. Furthermore, in some cases modules



Supplementary Figure 3.8: CS subtype identification by hierarchical clustering. For each of the training sets used to construct the five consensus sets (Table 3.1, main text) and for each of the IGLs S, H and P, four hierarchical clusterings were performed, labeled CS, CS+, CS++ and All (indicated on the x-axis for each panel). These respectively represent clusterings on the CS samples and three supersets of the consensus set. CS+: all samples for which PAM50 and all three SCMs are concordant, i.e. samples for which the St. Gallen criteria were left out of the CS inclusion criteria; CS++: all samples for which all three SCMs were concordant, i.e. samples for which the St. Gallen and the PAM50 CS inclusion criteria were not taken into account; All: the complete training set, i.e. when all CS inclusion criteria were dropped. Depicted are concordance (cc and kappa statistics) values between subtype assignments based on hierarchical clustering and the CS subtype labels. For a given set of samples concordance measures were always calculated on the CS samples only. The *intrinsic.cluster.predict* function from the *genefu* package was used to build a dendrogram (correlation distance, average linkage) and cut the dendrogram so as to obtain four clusters with a minimum of five samples per cluster (Haibe-Kains *et al.*, 2012). Concordance between the cluster labels and the consensus set subtype labels was determined by mapping clusters to a subtype label using the *matchClasses* function (method="exact") from the R package *e1071*. This function computes all possible permutations between rows and columns of the confusion matrix between two vectors of labels and selects the mapping such that as many cases as possible are in a matched pair. See Supplementary Table 3.12 for a detailed numerical summary.

were only weakly bimodal (BMI$\geq$1.1) or even not bimodal at all (BMI<1.1), in particular for the HER2-related module of Desmedt. Even though the module scores are not always strongly bimodal, the results provide solid ground for fitting the mixture models and cutoff values associated with SCM- and STG-based predictors.

Concordance of CS-based predictors on consensus sets

An important distinction between our approach and previous subtyping efforts is that our CS-based predictors were specifically designed to be highly concordant at the individual sample level. We first investigated the resubstitution performance, i.e. the ability of a CS-based predictor to correctly predict the subtype labels of the CS samples on which it was constructed. As expected, the resubstitution performance showed almost perfect levels of overall and

Subset	cc (all, %)	κ (all)	κ (basal)	κ (HER2)	κ (lumA)	κ (lumB)
CS	96.23	0.946	1.000	0.950	0.949	0.896
CS+	92.59	0.896	1.000	0.898	0.844	0.786
CS++	77.45	0.674	1.000	0.824	0.620	0.196
All	72.84	0.610	0.963	0.500	0.659	0.207

Supplementary Table 3.12: CS subtype identification by hierarchical clustering. Numerical details of Supplementary Figure 3.8: median percentage of concordant samples (cc) and median kappa statistics.

subtype-specific concordance (Supplementary Table 3.14).
A prerequisite for concordance over large validation cohorts is that predictors view each others training data in a consistent way. We, therefore, also considered the 'internal CS' validation performance, i.e. the ability of a CS-based predictor to predict the labels of all 812 CS samples, minus its own consensus training samples. Also in terms of internal CS validation performance, the CS-based predictors showed almost perfect levels of overall and subtype-specific concordance. The SCM.cs predictors showed the strongest levels of concordance (median κ=0.966, median cc=97.54%, Supplementary Table 3.15), closely followed by the SSP.cs predictors (median κ=0.940, median cc=95.66%), with equally strong subtype-specific levels of concordance. These results demonstrate that CS-based predictors are highly concordant on the individual sample level on training data.

	ER HK	ER D	ER W	HER2 HK	HER2 D	HER2 W	PGR	Proliferation	AURKA HK	AURKA D	AURKA W
Bos	2.45	2.26	2.09	1.76	1.28	2.26	1.97	1.40	1.08	1.24	1.36
expO	3.11	1.94	1.94	1.40	1.14	1.72	1.71	1.78	1.65	1.57	1.52
Guedj	2.87	1.91	1.90	1.24	0.86	1.67	1.95	1.79	1.71	1.64	1.61
Li	3.63	2.39	2.22	1.16	1.09	1.52	1.93	1.86	1.61	1.68	1.64
Sabatier	2.90	2.55	2.62	1.44	0.94	1.53	1.98	1.87	1.52	1.70	1.63
BMI (median)	2.90	2.26	2.09	1.40	1.09	1.67	1.95	1.79	1.61	1.64	1.61
Nr. BMI ≥ 1.1	5	5	5	5	2	5	5	5	4	5	5
Nr. BMI ≥ 1.5	5	5	5	1	0	5	5	4	4	4	4

Supplementary Table 3.13: Bimodality indices (BMI) of individual modules on consensus sets. Wang *et al.* (2009) characterize a distribution as being bimodal if BMI $\geq$ 1.1 and strongly bimodal if BMI ≥ 1.5. The first row indicates the various modules used to measure ER, HER2, PGR and proliferation (Section 3.7.1). Proliferation was measured by the AURKA proliferation modules by Haibe-Kains *et al.* (2012) (HK), Desmedt *et al.* (2008) (D) and Wirapati *et al.* (2008) (W) and the proliferation module (Proliferation) described in Supplementary Table 3.11. BMI values are listed for each consensus set. The last three rows provide the median BMI value over all five consensus sets, the number of times the module was bimodal and the number of times the module was strongly bimodal, respectively.

Subset	cc (all, %)	κ (all)	κ (basal)	κ (HER2)	κ (lumA)	κ (lumB)
All	98.80	0.983	1.000	1.000	0.991	0.983
SCM.cs	99.57	0.994	1.000	1.000	1.000	1.000
SSP.cs	97.65	0.967	0.945	0.987	0.983	0.954
SCM.cs HK	99.57	0.994	1.000	1.000	1.000	0.991
SCM.cs D	99.06	0.987	1.000	1.000	1.000	0.982
SCM.cs W	100.0	1.000	1.000	1.000	1.000	1.000
SSP.cs S	95.68	0.939	0.945	0.987	0.920	0.904
SSP.cs H	97.65	0.967	0.927	0.987	0.991	0.954
SSP.cs P	98.59	0.980	0.962	0.983	0.991	0.985

Supplementary Table 3.14: Resubstitution performance of CS-based predictors. Median percentage of concordant samples (cc) and median kappa statistics for CS-based predictors used to predict the subtype labels of their own consensus training set, i.e. to predict the associated CS labels. *Subset*: indicates the set of CS-based predictors over which the results were computed. Note that we report median values, it may therefore happen that for each individual subtype the median kappa statistic is equal to 1 but the overall median is not (2^{nd} row).

Subset	cc (all, %)	κ (all)	κ (basal)	κ (HER2)	κ (lumA)	κ (lumB)
All	96.91	0.957	0.948	0.990	0.953	0.938
SCM.cs	97.54	0.966	0.991	0.996	0.951	0.948
SSP.cs	95.66	0.940	0.931	0.983	0.956	0.902
SCM.cs HK	97.55	0.966	1.000	0.997	0.949	0.941
SCM.cs D	96.99	0.958	0.945	0.996	0.943	0.937
SCM.cs W	98.44	0.978	0.991	0.996	0.967	0.959
SSP.cs S	94.63	0.926	0.933	0.988	0.887	0.870
SSP.cs H	96.77	0.955	0.882	0.984	0.971	0.932
SSP.cs P	97.55	0.966	0.955	0.972	0.970	0.960

Supplementary Table 3.15: Internal CS validation performance of CS-based predictors. Median percentage of concordant samples (cc) and median kappa statistics for CS-based predictors used to predict the subtype labels of the union of all 812 CS samples, minus its own consensus training samples. *Subset*: indicates the set of CS-based predictors over which the results were computed.

3.7.3 Gene expression data

For the construction and evaluation of the consensus set-driven subtype predictors only high-quality Affymetrix arrays were used. This section gives a detailed description of the normalization and quality control (QC) stages used to process and filter these hybridizations. All analyses were performed using R/Bioconductor packages.

Normalization

In order to make the expression data as comparable as possible, we (re)normalized the Affymetrix datasets by a recently introduced modified version of the RMA methodology, known as frozen RMA (fRMA) (McCall *et al.*, 2010). This methodology allows one to normalize the intensity data of different arrays individually or in small batches and then combine the data for analysis. In particular, estimates of probe-specific effects and variances are pre-computed and frozen (McCall *et al.*, 2010). Another important distinction between default RMA and fRMA is the estimation of the reference distribution. In fRMA the reference distribution is not estimated from the data itself, but a pre-computed reference distribution is employed. Frozen RMA has the same logistical advantage as single chip models, in that it enables normalizing arrays one by one, while still having the benefits of a multi-chip normalization scheme. Our Affymetrix compendium involved two distinct array designs, i.e. hgu133plus2 and hgu133a arrays. We only considered the 22,215 probesets these designs have in common, which represent all non-control probesets present on the hgu133a platform. In order to utilize the common probesets,

the hgu133plus2 arrays were first converted to the hgu133a platform using the function *convertPlatform* from the *frma* package. We then masked all control probesets in the arrays and in the *hgu133afrmavecs* object containing the frozen parameters, resulting in the desired 22,215 probesets. In this way all Affymetrix arrays could be normalised using a single reference distribution, i.e. the Affymetrix hgu133a reference distribution, as constructed by McCall et al. based on 1,000 samples originating from 200 distinct studies (McCall and Irizarry, 2011). We ran *frma* in robust weighted average mode (McCall *et al.*, 2010).
Frozen RMA mainly addresses batch effects at probe level. fRMA-normalized data may therefore still contain batch effects at probeset level. Our Affymetrix compendium indeed showed clear evidence of systematic technical variation between arrays from different chip designs after fRMA (Supplementary Figure 3.9). This effect was removed via a robust scaling step (Materials and Methods, main text). A drawback of our approach is the loss of some hgu133plus2 probesets that are part of the gene list of certain subtype predictors. Some of these are Affymetrix control probesets which, interestingly, are included in the PAM50 gene list.

Quality control

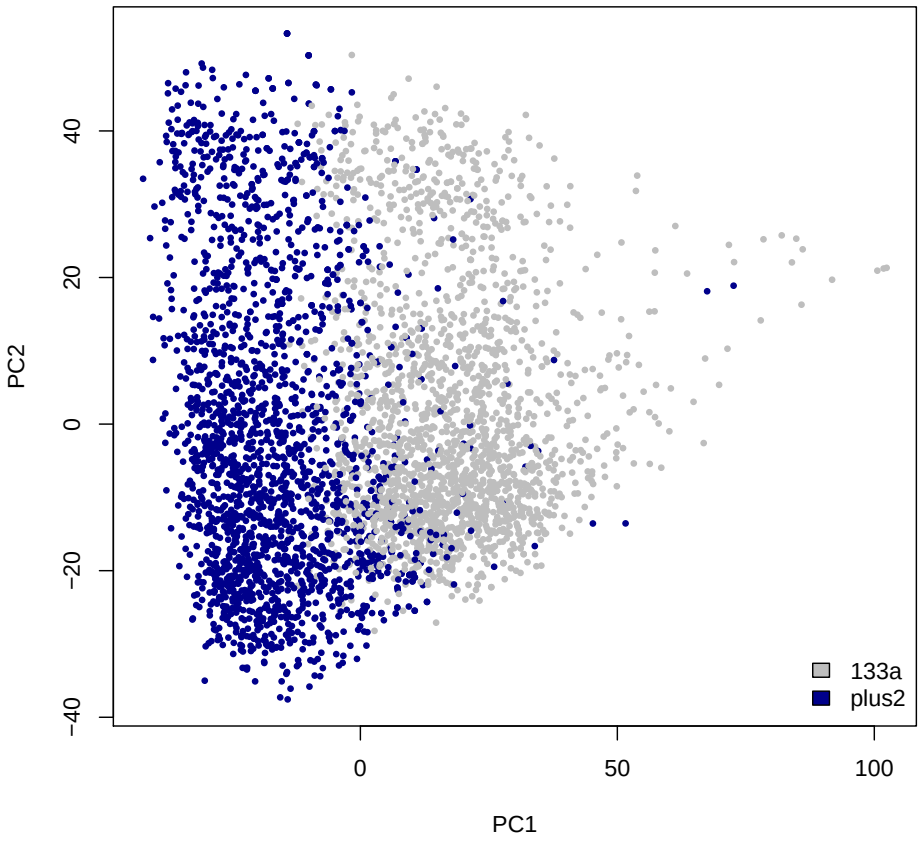
Poor hybridizations can have a negative impact on performance (Kauffmann and Huber, 2010). As we used datasets related to a substantial collection of high-quality publications, one may reasonably expect these hybridizations had passed quality control. However, after a preliminary QC inspection a sizable number of arrays appeared to be problematic for one or more well established QC control indicators. Supplementary Figure 3.10 provides several examples of problematic arrays encountered in our compendium. To ensure all hybridizations were of sufficient quality, an extensive QC analysis was performed aimed at identifying hybridizations that consistently showed indications of poor quality, either before or after normalization. The QC protocol we followed was based on six QC indicators: $Q = \{$RLE, NUSE, heatmap, boxplot, MA-plot, GNUSE$\}$. The first five represent well established QC indicators (Kauffmann and Huber, 2010). The GNUSE statistic was recently introduced by McCall *et al.* (2011) and is an fRMA-based single chip alternative to the multi-chip NUSE QC statistic (Bolstad, 2004). The NUSE, GNUSE and RLE QC indicators provide diagnostic information before normalization, while the remaining indicators provide information after normalization. All QC statistics with the exception of GNUSE were computed using the *arrayQualityMetrics* package, while GNUSE values were computed using the *frma* package. For a given QC indicator q and array i we used *arrayQualityMetrics* to obtain a series of QC scores and thresholds by repeatedly analyzing array i in the presence

of B randomly selected arrays from the same dataset. Higher scores reflect arrays of potentially poor quality, while scores higher than the threshold are considered outlier arrays. For a given array i and QC indicator $q \in Q$, let $S_{i,r}^q$ and τ_r^q be the QC score and threshold, respectively, as determined by *arrayQualityMetrics* at repeat r. Then, an array was rejected if it was considered an outlier in at least half of the QC repeats in which it was included. That is, array i was rejected based on QC if there exists a $q' \in Q$ for which we have

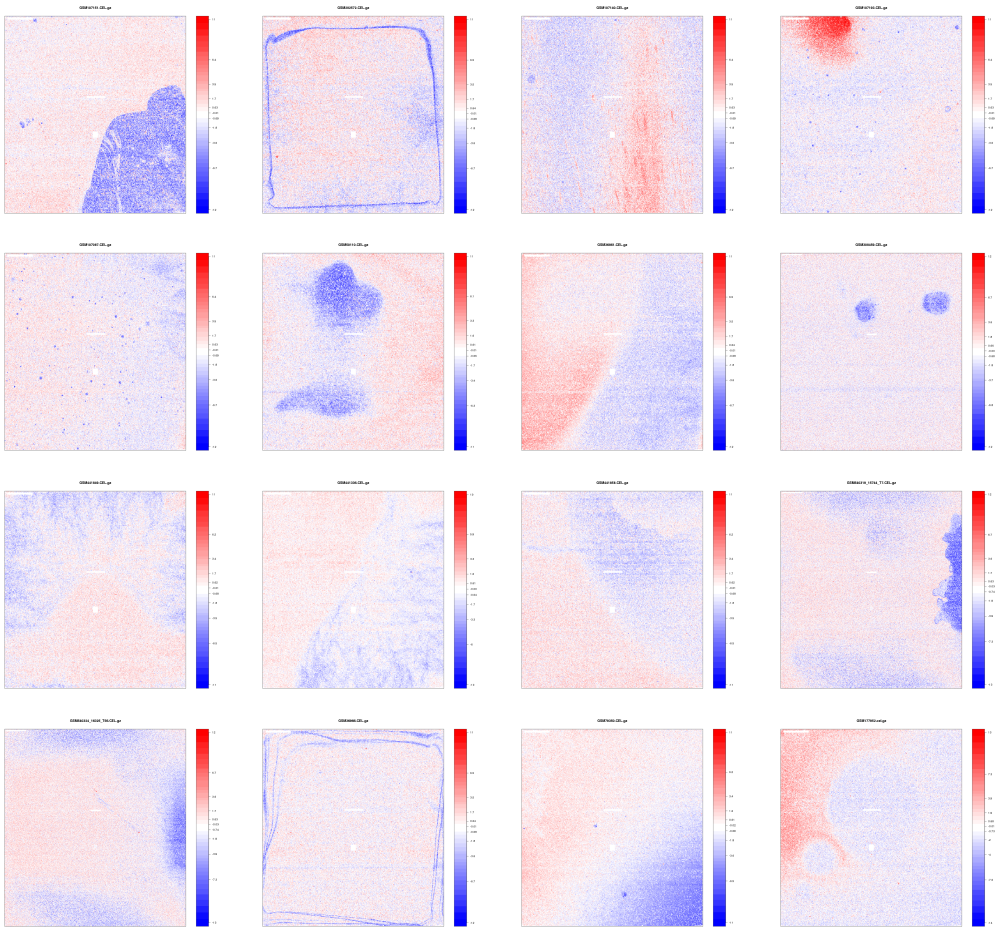
$$\sum_{r=1}^{R} I_{i,r}^{q'} \geq R/2$$

where $I_{i,r}^q$ is an indicator variable that equals 1 if $S_{i,r}^q > \tau_r^q$ and 0 otherwise and R is the number of repeats.

We ran the complete QC protocol on all 4,227 Affymetrix hybridizations part of our compendium. Arrays from different datasets and array designs were processed separately, with a QC batch size of $B = 30$ and $R = 10$ repeats. Hence, for each array and QC indicator we obtained 10 QC scores. In total 7.55% of the arrays (319 out of 4,227) were removed based on QC; 250 arrays (5.91%) showed consistent indications of poor quality prior to normalization and 182 (4.31%) after normalization; 2.67% (113 out of 4,227) of the arrays considered showed consistent indications of poor quality both before and after normalization. Supplementary Table 3.1 provides an overview of the QC results per dataset. For some datasets additional information was available on the processing groups (Leek *et al.*, 2010), e.g. the research institute in which the hybridizations were performed. In those instances QC batches were confined to include arrays from the same processing group only, even if this implied a batch size smaller than $B = 30$.



Supplementary Figure 3.9: Principal component analysis of fRMA-normalized data (combined hgu133plus2 and hgu133a compendium). Principal component (PC) analysis plot of the fRMA-normalized expression data from our Affymetrix compendium. Expression data originated from two chip designs, i.e. hgu133plus2 and hgu133a. In order to reduce systematic technical variation we used the frozen RMA methodology in which both array designs were normalized via a single reference distribution. A set of 3,400 genes related to breast cancer subtyping was used to estimate the principal components. This set corresponds to the union of all genes contained in the gene lists of the classic SSPs, classic SCMs and the CIT subtyping scheme of Guedj *et al.* (2011), for which probesets are present on the Affymetrix hgu133a design.



Supplementary Figure 3.10: Chip pseudo-images for 16 examples of arrays with consistent indications of poor quality. Details are provided in Supplementary Table 3.16.

x	y	ID	Dataset	Chip	GSM
1	1	771	Pawitan	hgu133a	GSM107151
1	2	1051	Schmidt	hgu133a	GSM282572
1	3	760	Pawitan	hgu133a	GSM107140
1	4	813	Pawitan	hgu133a	GSM107193.
2	1	708	Pawitan	hgu133a	GSM107087
2	2	670	MSK	hgu133a	GSM50110
2	3	1813	Wang	hgu133a	GSM36861
2	4	2343	Bos	hgu133plus2	GSM308459
3	1	415	Miller	hgu133a	GSM79350
3	2	1648	Symmans (II)	hgu133a	GSM441336
3	3	1564	Symmans (I)	hgu133a	GSM441858
3	4	4421	Sabatier	hgu133plus2	GSM540319_15744_T7
4	1	4426	Sabatier	hgu133plus2	GSM540324_16325_T56
4	2	1845	Wang	hgu133a	GSM36966
4	3	1218	Shi	hgu133a	GSM505494
4	4	163	Desmedt	hgu133a	GSM177952

Supplementary Table 3.16: Details on the 16 poor quality arrays from Supplementary Figure 3.10. x, y: coordinates of the examples, e.g. top left chip pseudo-image: x = 1, y = 1, bottom right: x = 4, y = 4; Chip: array design, GSM: accession number in GEO (Edgar *et al.*, 2002).

CHAPTER 4

EVALUATION STRATEGIES FOR SUBTYPE-SPECIFIC BREAST CANCER EVENT PREDICTION

4.1 Abstract

In recent years increasing evidence appeared that breast cancer may not constitute a single disease at a molecular level, but comprises a heterogeneous set of subtypes. This suggests that instead of building a single monolithic predictor, better predictors might be constructed that solely target samples of a designated subtype, which are believed to represent more homogeneous sets of samples. An unavoidable drawback of developing subtype-specific predictors, however, is that a stratification by subtype drastically reduces the number of samples available for their construction. As numerous studies have indicated sample size to be an important factor in predictor construction, it is therefore questionable whether the potential benefit of subtyping can outweigh the drawback of a severe loss in sample size. Factors like unequal class distributions and differences in the number of samples per subtype, further complicate comparisons. We present a novel experimental protocol that facilitates a comprehensive comparison between subtype-specific predictors and predictors that do not take subtype information into account. Emphasis lies on careful control of sample size as well as class and subtype distributions. The methodology is applied to a large breast cancer compendium involving over 1500 arrays, using a state-of-the-art subtyping scheme. We show that the resulting subtype-specific predictors outperform those that do not take subtype information into account, especially when taking sample size considerations into account[1].

[1]This work was published as: HMJ Sontrop, WFJ Verhaegh, MJT Reinders, PD Moerland (2011). An evaluation protocol for subtype-specific breast cancer event prediction. PLoS ONE, 6(7):e21681.

4.2 Background

Breast cancer *event prediction* is an important yet challenging classification problem in which one attempts to predict whether a certain type of event will happen within a given time frame or not, e.g. whether a breast tumor will metastasize or not, based on gene expression data obtained from microarrays. A well-known example of such a predictor is the 70-gene signature by van't Veer *et al.* (2002). In recent years increasing evidence appeared implying that breast cancer may not constitute a single disease at the molecular level, but that breast cancers comprise a diverse and heterogeneous set of diseases (Weigelt *et al.*, 2010d).
Various breast cancer subtyping schemes have been proposed, mostly inspired by the *intrinsic gene list* approach from the landmark publication by Perou *et al.* (2000). The latter introduced a breast cancer subtype taxonomy that classifies breast cancers as either luminal A (lumA), luminal B (lumB), basal, HER2 or normal-like, based on hierarchical clustering. A more recent example is a subtyping scheme based on a biology-inspired module-driven approach (Desmedt *et al.*, 2008), that identifies the subtypes lumA, lumB, basal, and HER2 through model-based clustering. The precise definition of the subtypes themselves and of a standardized geneset to classify samples to a specific subtype is still subject of debate. Several studies indicated stability issues with the intrinsic gene list approach (Kapp *et al.*, 2006; Pusztai *et al.*, 2006; Weigelt *et al.*, 2010a). Furthermore, doubts have been casted on the existence of the normal-like tumours as a genuine breast cancer subtype (Parker *et al.*, 2009). Despite this debate, it is widely accepted that over large sample sets breast cancer subtypes are associated with a difference in survival time. This suggests that instead of using a single monolithic predictor, better prognostic predictors might be constructed that solely target samples of a designated subtype. However, only few studies couple subtyping directly to breast cancer event prediction (Haibe-Kains *et al.*, 2010; Parker *et al.*, 2009; Wang *et al.*, 2005). In this chapter we address the question whether predictors targeting a specific subtype, referred to as *typed* predictors, can outperform *untyped* predictors that do not take subtype into account. The main contribution of this work is the definition of a novel experimental protocol which explicitly addresses three main problems of such a comparison, i.e. subtype definition, sample size, and class imbalance.

Subtype definition

In this chapter we are interested in the possibilities of improving microarray breast cancer event prediction by exploiting subtype information. A core ingredient of our protocol is the construction of a sequence of subtype-specific

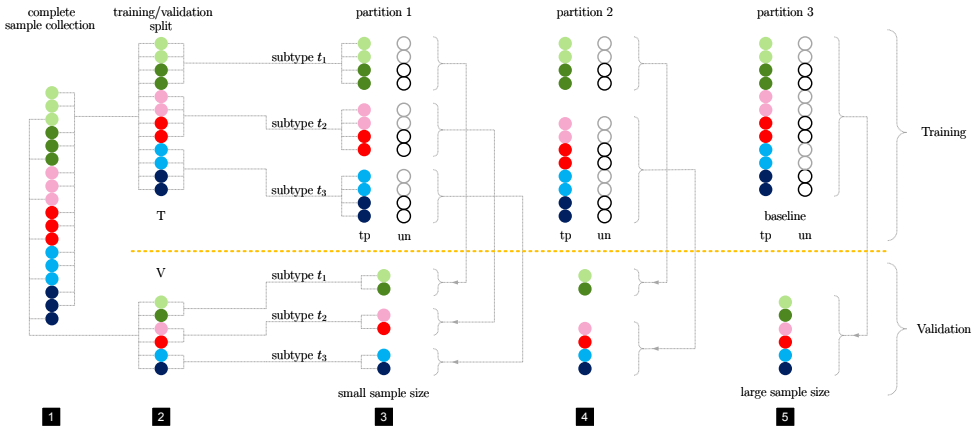


Figure 4.1: Conceptual overview of the stratification protocol. 1) toy sample set, comprised of three subtypes (blue, red and green), lighter (darker) shades indicate positive (negative) cases. 2) stratified split (by class label and subtype) of the data into a training set T and a validation set V. For each set separately various partitions are created. The yellow dashed line illustrates the strict separation of training (top) and validation (bottom) parts. 3) the most refined partition involves a single subtype per part. The typed version (tp) partitions T by parts stratified by class label and subtype. The untyped (un) counterpart involves parts stratified by class label only, however, each untyped part involves an identical number of positive and negative training samples as its typed counterpart. Here lighter (darker) open circles represent positive (negative) cases. Alternative partitions can be constructed by pooling some or all of the initial parts, as depicted in 4) and 5). On each training part a separate predictor is constructed, which is evaluated on a specific set of validation samples. Note that paired typed and untyped predictors are evaluated on the same set of validation samples. 5) presents a special case for which typed and untyped training sets are identical and equal the overall training set T. This set is used to construct the baseline predictor. The untyped predictors associated with partitions 1 and 2 represent down-scaled versions of the baseline and serve to assess the influence of sample size.

predictors that via systematic pooling steps gradually transform into an untyped baseline predictor. A conceptual overview of the stratification of subtypes is provided by Figure 4.1. From the application of a given subtyping scheme, e.g. the module-based approach of (Desmedt *et al.*, 2008), each sample is associated with a specific subtype. These subtype labels are subsequently used to construct various partitions of the available data. For each part of a partition a separate predictor is constructed, which targets a specific subset of samples. The most refined partition contains one subtype per part. From this partition a sequence of alternative partitions is created by systematic pooling of individual parts. Ultimately, this leads to a partition with a single part. The performance of this partition serves as a natural baseline as its associated predictor is essentially untyped and is constructed on the largest sample set available, which simultaneously represents the most heterogenous

	lumA	lumB	basal	HER2	D
N_s	273 (41.2)	216 (32.8)	100 (15.1)	74 (11.2)	663 (100)
P_s	42 (18.3)	94 (41.0)	57 (24.9)	36 (15.7)	229 (100)
total	315 (35.3)	310 (34.8)	157 (17.6)	110 (12.3)	892 (100)
ratio	6.5	2.8	1.8	2.1	2.9

Table 4.1: Distribution of class labels and subtypes for the 892 samples with a proper class label. N_s and P_s denote the number of negative (good prognosis) and positive (poor prognosis) cases of for each subtype s, *total* and *ratio* represent the sum and ratio of N_s and P_s, respectively. Entries in brackets indicate percentages w.r.t. the entire compendium (column D).

set w.r.t. the selected subtyping scheme. For a given partition, of interest are the performance per part, as well as the overall performance associated with it, that is, the performance as evaluated over all available samples. We note that, even though the set of subtypes used to construct partitions is of great interest, its precise makeup is of a lesser concern in this chapter, as we are mainly concerned in setting up a proper comparison between partitions.

Sample size

The sample size problem manifests itself in different ways. Firstly, stratification by subtype drastically reduces the size of the sample set available for the construction of typed predictors (Figure 4.1). As numerous studies have shown that a larger sample size leads to better performance (Michiels *et al.*, 2005; van Vliet *et al.*, 2008; Kim, 2009) it is therefore non-trivial if the potential benefit of subtyping can outweigh a severe loss in sample size. Secondly, differences in sample size per subtype also complicate the comparison between typed predictors. This imbalance is clearly illustrated by the application of a state of the art model-based subtyping scheme (Desmedt *et al.*, 2008) to a compendium of 892 breast cancer samples (Table 4.1) used in this chapter. Our experimental protocol strongly controls these sample size effects to enable a systematic comparison of typed and untyped predictors.

Class imbalance

Imbalance with respect to the class label distribution is another important characteristic of many cancer related datasets. Also in our breast cancer compendium the positive class, i.e. the poor prognosis group, is much smaller than the negative class, i.e. the good prognosis group (Table 4.1, column D). Such imbalance often negatively affects the performance of a predictor for the minority class. The literature offers several solutions for the class imbalance problem. Popular approaches are to either undersample the majority class,

to oversample the minority class, or to adapt the cost structure (Blagus and Lusa, 2010; He and Garcia, 2009). This is especially important in a subtyping setting where a proper comparison of predictors is affected by a class imbalance inherent to the subtyping itself. Note that, if the subtype has a profound impact on the survival rate, we expect distinct subtypes to be associated with different negative to positive class ratios. In our compendium, we see that this is indeed the case (Table 4.1). Comparisons between predictor performances using frequently adopted performance measures like accuracy, positive and negative predicted value, can easily be obscured by a difference in the class ratio. For these reasons, proper balancing is essential.

This chapter presents a novel experimental evaluation protocol that highly facilitates the comparison between typed and untyped predictors, in which sample size as well as class and subtype distributions are controlled and by which their individual contributions can be properly studied. In order to facilitate a proper comparison, besides working with the complete (unbalanced) compendium, we also consider performance on a set of *balanced* compendia which have the same sample size and negative-positive class ratio for each subtype and are obtained via undersampling of the majority class. Although here applied to microarray breast cancer event prediction, the methodology is also applicable to other types of diseases or data obtained by alternative measurement techniques.

4.3 Materials and Methods

The following sections present a novel predictor construction and evaluation protocol to investigate the potential of typed prediction and its relation to sample size. The upcoming sections present a detailed formal description of the various individual steps involved in the predictor construction and evaluation protocol. A bird's-eye view of the complete protocol is provided by Figure 4.4 on page 131. The protocol produces a sequence of predictors that via systematic pooling steps gradually transform into an untyped baseline predictor. As appropriate choices for a prediction rule, ranking, subtyping strategy, and performance measure are domain-specific, for the moment we assume they are given.

4.3.1 Partitioning scheme

Let D denote the set of all available samples with proper event data, that are associated with a set of n *elementary* subtypes $S_e = \{t_1, \cdots, t_n\}$. The elementary subtypes form the most obvious candidates to consider for typed

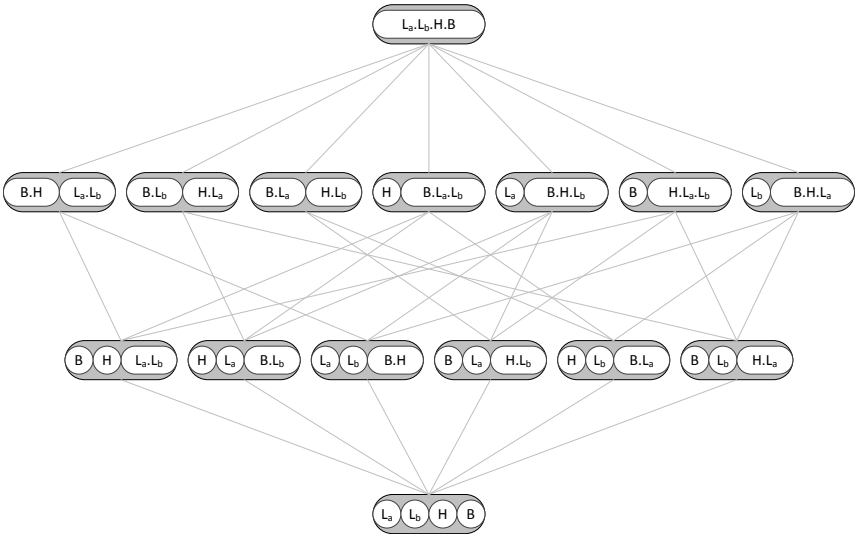


Figure 4.2: Partitioning scheme. The Hasse diagram depicts all possible partitions (grey ovals) w.r.t. an example breast cancer subtype set $S_e = \{L_a, L_b, H, B\}$, representing the subtypes lumA, lumB, HER2, and basal, respectively. White ovals indicate parts. The lines represent a move from one partition to another by either merging two parts (bottom to top) or splitting one part into two parts (top to bottom). The top layer depicts the coarsest partition in which all elementary types have been pooled into a single part, making it essentially untyped. The bottom layer represents the most refined partition, i.e. one part for each elementary subtype. For each distinct part a separate predictor is constructed. The partition in the top layer is used for baseline predictor construction. In this example $B_n = 15$, $|S_e| = 4$, $|S| = 15$ and $|S_c| = 11$.

prediction. In this case one would partition the available sample set D into exactly n parts. Less refined partitions, however, can be considered by pooling members of several elementary subtypes, ultimately leading to a single part, that is essentially untyped. Let S denote the collection of distinct parts over all partitions, that is, the powerset of S_e minus the empty set with cardinality $|S| = 2^n - 1$. We will refer to the set $S_c = S - S_e$ as the set of *compound* subtypes, the members of which are comprised of several of the elementary subtypes. In general, the number of distinct partitions is given by the n^{th} Bell number (Rota, 1964), denoted by B_n, where n represents the number of elementary subtypes. The complete set of partitions can be conveniently arranged into a Hasse diagram, see Figure 4.2, which shows an example for $n = 4$ elementary subtypes.

4.3.2 Evaluation protocol and predictor construction

In essence our evaluation protocol can be seen as an extension of the protocol proposed by Wessels *et al.* (2005). Our protocol consists of a repeated stratified

cross-validation scheme for the typed predictors, after which we deliberately randomize the corresponding training sets w.r.t. subtype distribution, in order to obtain results for the untyped predictors. Below we give a formal description of the protocol.

Notation

Let P_s and N_s denote the sets of positive and negative samples of subtype s. For each $s \in S_e$ we divide the corresponding sets P_s and N_s into K_{out} folds of approximately the same size. Let F denote the set of all folds, with $|F| = K_{\text{out}}$, let $P_{s,f}$ ($N_{s,f}$) denote fold f of P_s (N_s) and let $P_{s,\text{-}f}$ ($N_{s,\text{-}f}$) denote the union of all folds but fold f. Now we can define the training and validation sets for typed and untyped predictors. A detailed toy example clarifying the sets as defined in the following two subsections is provided by Figure 4.3.

Typed sets

For each elementary subtype $s \in S_e$ and fold $f \in F$ we construct a typed training set $T^{\text{tp}}_{s,f} = P_{s,\text{-}f} \cup N_{s,\text{-}f}$ and a validation set $V_{s,f} = P_{s,f} \cup N_{s,f}$. Furthermore, for each compound subtype and fold we pool the training and validation sets of the subtypes that comprise it, that is, for compound subtype $s' \in S_c$ consisting of the elementary subtypes $S' \subseteq S_e$ we have $T^{\text{tp}}_{s',f} = \bigcup_{s \in S'} T^{\text{tp}}_{s,f}$ and $V_{s',f} = \bigcup_{s \in S'} V_{s,f}$.

Untyped sets

In order to construct untyped counterparts of the typed training sets let $P_{\text{-}f} = \bigcup_{s \in S_e} P_{s,\text{-}f}$ and $N_{\text{-}f} = \bigcup_{s \in S_e} N_{s,\text{-}f}$. For each elementary subtype $s \in S_e$ and fold $f \in F$ we create the sets $P^{\text{un}}_{s,\text{-}f}$ and $N^{\text{un}}_{s,\text{-}f}$ by randomly drawing *without replacement* $|P_{s,\text{-}f}|$ positive and $|N_{s,\text{-}f}|$ negative samples from the sets $P_{\text{-}f}$ and $N_{\text{-}f}$, respectively. Analogously to the typed scenario, for each elementary subtype $s \in S_e$ and fold $f \in F$ we next construct an untyped training set $T^{\text{un}}_{s,f} = P^{\text{un}}_{s,\text{-}f} \cup N^{\text{un}}_{s,\text{-}f}$, which has the same negative to positive ratio as $T^{\text{tp}}_{s,f}$. Finally, for each compound subtype and fold we again pool the corresponding training sets of the elementary subtypes that comprise it, that is, for compound subtype $s' \in S_c$ consisting of the elementary subtypes $S' \subseteq S_e$ we have $T^{\text{un}}_{s',f} = \bigcup_{s \in S'} T^{\text{un}}_{s,f}$. Typed and untyped predictors are paired and their performance is evaluated on the same validation set.

Baseline

Note that the only partition for which typed and untyped sets are identical is the partition in which all elementary subtypes have been pooled into one part.

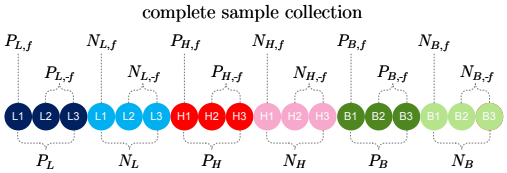
In this case typed and untyped predictors for each fold $f \in F$ are associated with the same training set $T_f = \bigcup_{s \in S_e} T^{\text{tp}}_{s,f} = \bigcup_{s \in S_e} T^{\text{un}}_{s,f}$, with corresponding validation set $V_f = \bigcup_{s \in S_e} V_{s,f}$. We will refer to these predictors as *baseline predictors*.

Toy example visualizing the construction of typed and untyped set

Consider the balanced toy dataset depicted in Panel A) of Figure 4.3, which is an extension of the example depicted in Figure 4.1. The sample set is again comprised of three three elementary subtypes, $S_e = \{\{L\}, \{H\}, \{B\}\}$, representing for instance the subtypes luminal (blue), HER2 (red), and basal (green), respectively. Each elementary subtype consists of three positive (poor prognosis) cases, depicted by darker shades and three negative (good prognosis) cases, depicted by lighter shades. Instead of an individual sample (Figure 4.1), here each circle corresponds to multiple samples. Panel B) depicts the associated Hasse diagram w.r.t. the elementary subtype set S_e with five partitions (see also Figure 4.2). Panel C) presents an overview of the five typed partitions of the Hasse diagram in the context of a $K_{\text{out}} = 3$-fold cross-validation scheme. The example depicts the sets associated with a single fold. Validation sets are depicted at the left of the vertical dotted line, training sets on the right. Each part in a partition is depicted as a connected string of filled circles. For each training part a separate predictor is constructed. Partition names are given at the outer right, where a dot indicates pooling, and a vertical dash is used to separate parts. Finally, Panel D) depicts five untyped partitions for a single fold. The untyped training set for the most refined partition (#5) is constructed from the typed training set by randomly swapping light shaded training instances with each other and dark shaded instances with each other. This guarantees that the negative-positive class ratio is the same for typed and untyped sets. Coarser partitions (#1-4) are formed by combining parts according to the Hasse diagram of panel B. Note that for the coarsest partition (#1), typed and untyped training sets are identical. This set is used for the construction of the baseline predictor. Last, note that typed and untyped partitions are always associated with the same set of validation samples. Furthermore, training and validation samples are always strictly separated.

Figure 4.3 ***(facing page)*****: Stratification toy example**. For a detailed explanation, see the Section 4.3.2.

A

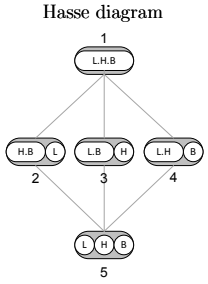


$S_e = \{\{L\}, \{H\}, \{B\}\}$

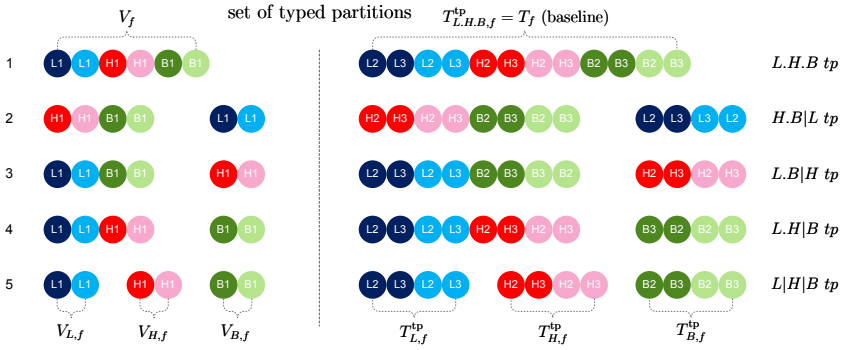
$S_c = \{\{L, H\}, \{L, B\}, \{H, B\}, \{L, H, B\}\}$

$S = S_e \cup S_c$

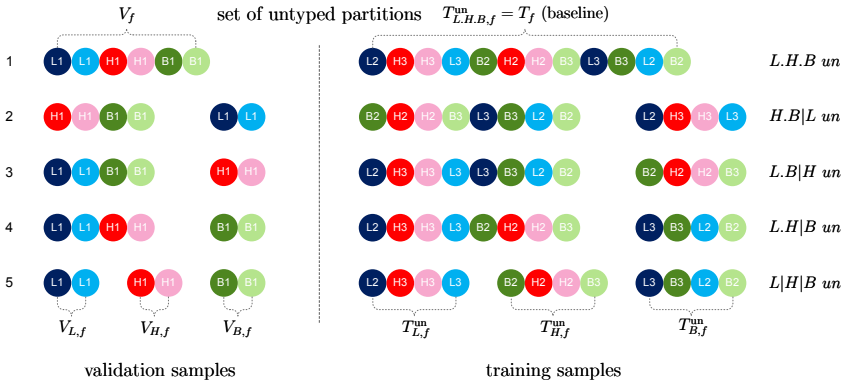
B



C



D



Training protocol

On every training set we invoke an identical training protocol, which is a mild adaptation of the protocol proposed by Wessels *et al.* (2005). Let T' denote the set of available training samples. In a first step we divide T' into K_{in} folds stratified w.r.t. class label and subtype. For each fold g we perform a ranking using the learning set $L_g = T'_{-g}$, after which we construct a sequence of $d_{\max}$ predictors C_d using the top $d \in \{1, 2, \ldots, d_{\max}\}$ ranked features on L_g. We then employ these predictors to predict the events corresponding to the evaluation set $E_g = T'_g$ and subsequently aggregate the results over all folds from which we construct a performance curve, which for a performance indicator of interest tells us the performance for a given number of features, up to $d_{\max}$. The previous training steps are repeated R_{in} times in order to construct an average performance curve which for a given set size reports the average performance over all repeats. We refer to this loop as the *inner loop* of our protocol.

Let μ^* denote the maximum value of the average performance curve and denote its standard deviation over R_{in} repeats by σ^*. Since larger signatures are often more robust (Sontrop *et al.*, 2009), we take the optimal number of features to be the largest integer $d^* \leq d_{\max}$ such that its associated training performance $p^* \geq (\mu^* - \sigma^*)$. Finally, we use the full training set T' to rank the available features and construct a predictor C_{d^*} using the top d^* ranked features on T' and conclude by returning p^*, d^*, as well as the trained predictor C_{d^*}. In addition to an optimized signature size d^*, a fixed size can be considered as well.

Performance evaluation

For each subtype $s \in S$ and for each fold $f \in F$ we invoke the training protocol on the typed and untyped training sets, $T^{\text{tp}}_{s,f}$ and $T^{\text{un}}_{s,f}$, and apply both of the resulting predictors to the same validation set $V_{s,f}$. Let $A^{\text{tp}}_{s,f}$ and $A^{\text{un}}_{s,f}$ denote the assignments made on this validation set by the typed and untyped predictors, respectively. For each subtype s we construct a subtype-specific performance indicator for the typed and untyped predictors by considering the aggregated assignments over all folds $A^{\text{tp}}_s = \bigcup_{f \in F} A^{\text{tp}}_{s,f}$ and $A^{\text{un}}_s = \bigcup_{f \in F} A^{\text{un}}_{s,f}$. Finally, for a given partition P we obtain an overall performance estimate for typed and untyped predictors by considering the aggregated assignments over all its parts $A^{\text{tp}} = \bigcup_{s \in P} A^{\text{tp}}_s$ and $A^{\text{un}} = \bigcup_{s \in P} A^{\text{un}}_s$, respectively. To compensate for sampling effects all previous steps are repeated R_{out} times, after which we average performance indicators over all repeats. We refer to this loop as the *outer loop*.

4.3.3 Performance measures

Class imbalance influences the choice of a suitable performance measure. Comparison of performance by the total accuracy rate has the disadvantage that a predictor that always guesses the majority class is associated with a high performance, while in fact it misclassifies the complete minority class. A more appropriate performance measure is the area under the ROC curve, which is insensitive to varying class proportions. Also the *balanced accuracy rate*, defined as the average of the sensitivity and specificity of the prediction rule, has been used in an imbalanced setting (Wessels *et al.*, 2005; Parker *et al.*, 2007; van Vliet *et al.*, 2008). This measure has the advantage that we can no longer achieve a high performance by sacrificing one class for another, as doing so results in a performance equal to that obtained by random guessing, i.e. a balanced accuracy rate of 50%.

Our main performance indicator is the area under the ROC curve (*auc*). We also report the balanced accuracy rate (*bar*) and the accuracy (*acc*). Since summarizing predictor performance on both classes in a single measure causes loss of information, we also report four other frequently used performance indicators that report performance for a proper subset of the samples: sensitivity (*sen*), specificity (*spc*), positive predictive value (*ppv*), and negative predictive value (*npv*). For a thorough overview of these and other performance indicators see Baldi *et al.* (2000).

4.3.4 Balanced compendia

Since the number of samples and the negative-positive class ratio differ considerably per subtype (Table 4.1), we constructed a set of balanced compendia that are properly stratified w.r.t. the class ratio. Note that the largest sample set that can be constructed with the same number of samples and the same ratio $q_s = |N_s|/|P_s|$ for all elementary subtypes can hold at most $m_N = \min\{|N_{t_1}|, \ldots, |N_{t_n}|\}$ negative samples and $m_P = \min\{|P_{t_1}|, \ldots, |P_{t_n}|\}$ positive samples. Therefore, in order to obtain a balanced compendium B, we randomly draw *without replacement* m_N negative samples from N_s and m_P positive samples from P_s for each elementary subtype $s \in S_e$. Let B_s denote the set of $m_P + m_N$ samples drawn for subtype $s \in S_e$, then $B = \bigcup_s B_s$. Since for most elementary subtypes the sampling can be done in multiple ways, we repeat the sub-sampling process R_{bal} times. Note that, compared to the unbalanced compendium D, the balanced compendia B are well controlled w.r.t. subtype distribution, sample size, and class distribution.

4.3.5 Schematic representation main evaluation protocol

Figure 4.4 presents a schematic representation of the main evaluation protocol as described above when applied to the toy dataset example of Figure 4.3. For clarity the figure depicts the scenario for a single fold f and depicts only two of the $B_n = 5$ partitions i.e. the coarsest (partition 1, Figure 4.3) and the most refined (partition 5, Figure 4.3). The former partition is associated with the baseline predictor, for which typed and untyped are identical, and involves steps 1, 4, 8, 11, and 14 of Figure 4.4. The second partition contains one part for each elementary subtype. Typed predictors involve steps 2, 5, 9, 12, and 15, while untyped predictors involve steps 3, 6, 10, 13, and 16.

4.3.6 Compendium construction

The compendium pools data of ten individual microarray datasets. All datasets were measured on the same platform (Affymetrix HG-U133A). This circumvents the need for cross-platform normalization, which can be challenging (Perou *et al.*, 2010). All raw expression data used is publicly available in the MIAME compliant databases Gene Expression Omnibus (GEO) (Edgar *et al.*, 2002) and ArrayExpress (Parkinson *et al.*, 2005) and can be found under the following accession numbers: GSE2034 (Wang *et al.*, 2005), GSE5327 (Yu *et al.*, 2007), GSE7390 (Desmedt *et al.*, 2007), GSE11121 (Schmidt *et al.*, 2008), GSE2603 (Minn *et al.*, 2005), GSE6532 (Loi *et al.*, 2007), GSE2990 (Sotiriou *et al.*, 2006), GSE3494 (Miller *et al.*, 2005), GSE1456 (Pawitan *et al.*, 2005), and E-TABM-158 (Chin *et al.*, 2006). All accession numbers represent GEO accession numbers, with exception of E-TABM-158 (Chin *et al.*, 2006), the expression data of which is stored at ArrayExpress. After removing duplicate entries and outlier arrays, detected using the *arrayQualityMetrics* package (Kauffmann *et al.*, 2009), 1539 unique hybridizations remained. Raw expression data was used to generate MAS5.0 expression estimates, using the *affy* package, scaled to a target intensity of 600. Prior to pooling expression data, the expression estimates were z-transformed for each study and each gene separately, as suggested in Yasrebi *et al.* (2009); Perou *et al.* (2010). For event prediction purposes, all class labels are solely based on a single type of survival data, being distant metastasis free survival (dmfs). *Poor prognosis* cases (PP) had an event, i.e. distant metastasis within five years, while the *good prognosis* cases (GP) did not have an event during follow-up, with a follow-up time of at least five years i.e. samples with an event after five years were removed. These stringent criteria led to the identification of 229 PP samples and 663 GP samples, yielding a total of 892 unique samples. A list of the individual CEL file identifiers is presented in Supporting Information file

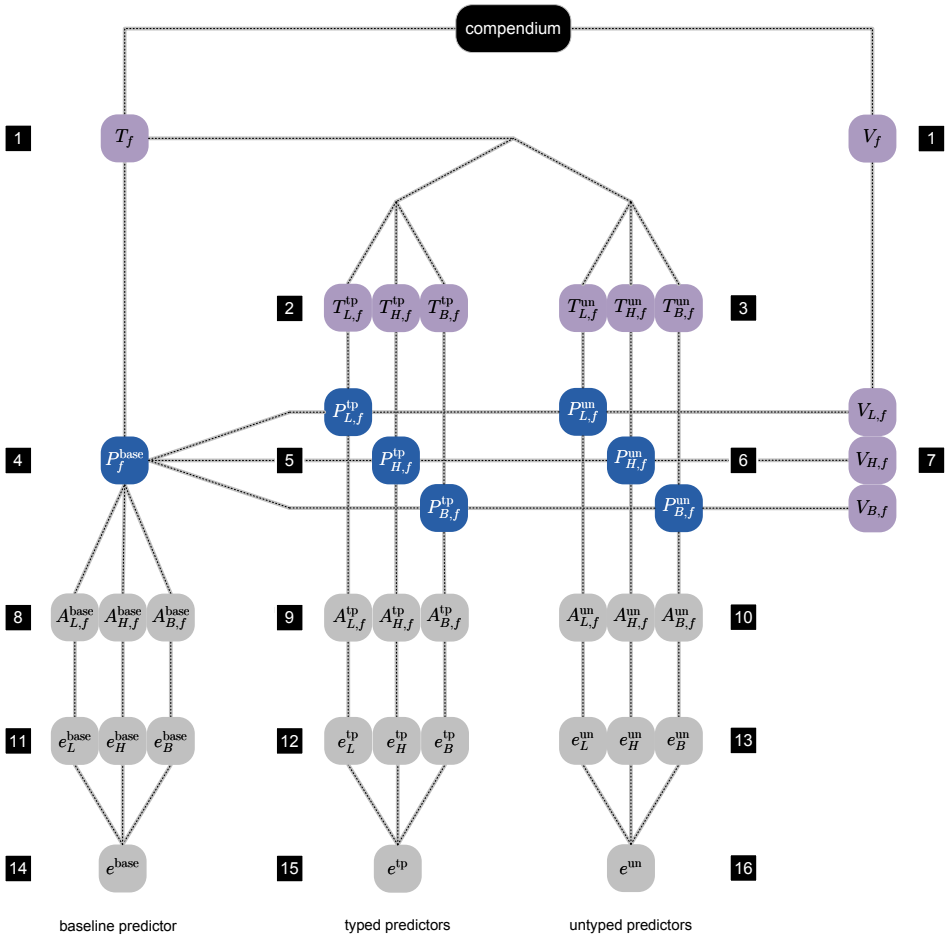


Figure 4.4: Bird's-eye view of evaluation protocol. 1) Stratified split w.r.t. class label and subtype of the complete data set in a training set T_f and a validation set V_f. 2) Construction of typed training sets $T^{\text{tp}}_{L,f}$, $T^{\text{tp}}_{H,f}$ and $T^{\text{tp}}_{B,f}$. 3) Construction of untyped training sets $T^{\text{un}}_{L,f}$, $T^{\text{un}}_{H,f}$ and $T^{\text{un}}_{B,f}$. 4) Baseline predictor construction. 5) Typed predictor construction. 6) Untyped predictor construction. 7) Stratification of validation set by subtype. 8) Invoke baseline predictor on validation samples. 9) Invoke typed predictors on associated validation samples. 10) Invoke matching untyped predictors on same validation sets. Steps 1-10 are repeated for all folds $f \in F$. 11-13) Subtype-specific performance estimation based on the aggregated event predictions (over all folds) per subtype, as made by the baseline (11), typed (12), and untyped (13) predictors. 14-16) Overall performance estimation based on the aggregated event predictions over all folds made by the baseline (14), typed (15), and untyped (16) predictors.

S1 online[2].

4.3.7 Subtyping scheme

Subtyping is based on the biology-inspired module-driven approach by Desmedt *et al.* (2008), that identifies the subtypes lumA, lumB, basal, and HER2 through model-based clustering. In contrast to the intrinsic gene list approach (Perou *et al.*, 2000), clustering is not performed on the expression data directly. Instead the expression values are first projected onto a lower dimensional space, in which each sample is represented by three *module scores* related to key biological processes strongly associated with breast cancer. The modules consist of an ER-related module, comprising 469 genes, a HER2-related module of 28 genes, and a proliferation-related module, referred to as AURKA, containing 229 genes. After transformation of the expression data to module scores, a Gaussian mixture model is fitted on the module data in order to determine the cluster membership of each sample. ER and HER2 module scores are used to infer the subtypes luminal, HER2, and basal, while the AURKA module is used to further subdivide the luminal group into a lumA and a lumB group.
In order to obtain the most likely subtype assignment for each sample, we estimated the subtype model on the set of all 1539 available samples. This resulted in 564 (36.8%), 543 (35.4%), 246 (17.6%) and 186 (16.1%) assignments to the subtype categories lumA, lumB, basal, and HER2, respectively. Table 4.1 presents an overview of these assignments for the set of 892 samples with properly defined class labels. The subtype distribution over the 892 sample set is similar to the subtype distribution over the complete compendium with 35.3%, 34.8%, 17.6%, and 12.3% belonging to the subtypes lumA, lumB, basal, and HER2, respectively ($P = 0.95$, Pearson's chi-square test). Subtyping was performed using the *genefu* package.

4.3.8 Balanced sets

From Table 4.1 it follows that in order to obtain a fully balanced compendium, we can select at most $m_N = 36$ negative and $m_P = 74$ positive cases for each $s \in S_e$, which in turn implies $|B_s| = 36 + 74 = 110$ and $|B| = 4 \times 110 = 440$.

Protocol implementation details

This chapter presents results over a set of $R_{\text{bal}} = 100$ balanced breast cancer compendia, and for an unbalanced compendium of 892 samples.

[2]See http://www.plosone.org/article/info%3Adoi%2F10.1371%2Fjournal.pone.0021681#s5 for all online Supporting Information files corresponding to this chapter.

For the inner loop we employed $K_{\text{in}} = 10$-fold cross-validation, with $R_{\text{in}} = 5$ repetitions. Predictors are based on the nearest centroid (NC) rule, which despite its simplicity often shows good performance. Furthermore, a NC is known to be reasonably noise tolerant (Wessels *et al.*, 2005). As a distance measure the cosine correlation distance was used. For each separate fold of the training set we first performed a filtering step, using the present/absent calls from the MAS5.0 procedure and only selected genes for which in at least one of the positive or negative sample groups the number of present calls was at least 70% (McClintick and Edenberg, 2006). The remaining features were ranked based on moderated-t statistics, as implemented in the *limma* package (Smyth, 2005). For predictor construction we considered average performance curves up to $d_{\max} = 200$ features, similar to van Vliet *et al.* (2008). Finally, in the outer loop we employed $K_{\text{out}} = 10$-fold cross-validation, with $R_{\text{out}} = 100$ repetitions. ROC curves were generated by using the difference between the distance of a sample to each of the centroids as a continuous criterion, on which a variable threshold was set.

4.4 Results

4.4.1 Improved auc and bar by typed prediction

Figure 4.5 depicts a condensed overview of overall performance corresponding to typed and untyped event predictors under various partitioning schemes, involving signatures based on the nearest centroid rule. Similar results were obtained using a signal-to-noise ratio ranking strategy, using 3-fold, 5-fold, and leave-one-out cross-validation instead of 10-fold cross-validation, or when using a more complex non-linear predictor (random forest (Breiman, 2001)), see Supporting Information file S2 online. A complete overview of the performance per subtype associated with Figure 4.5 is given in Supporting Information file S3 online.

Performance on balanced compendia

The left panel in Figure 4.5 shows that typed predictors generally obtain a higher overall performance than their untyped counterparts on balanced compendia. The typed *auc* and *bar* are consistently higher, sometimes quite substantially. Furthermore, we see that *auc* and *bar* are well correlated.
One of the more interesting partitions is the one that uses a single part for each elementary subtype, which is situated at the outer right in each panel and corresponds to the partition depicted at the bottom of the Hasse diagram (Figure 4.2). In this partition overall performance in the typed case is obtained by employing four distinct typed predictors, each targeting a different part of

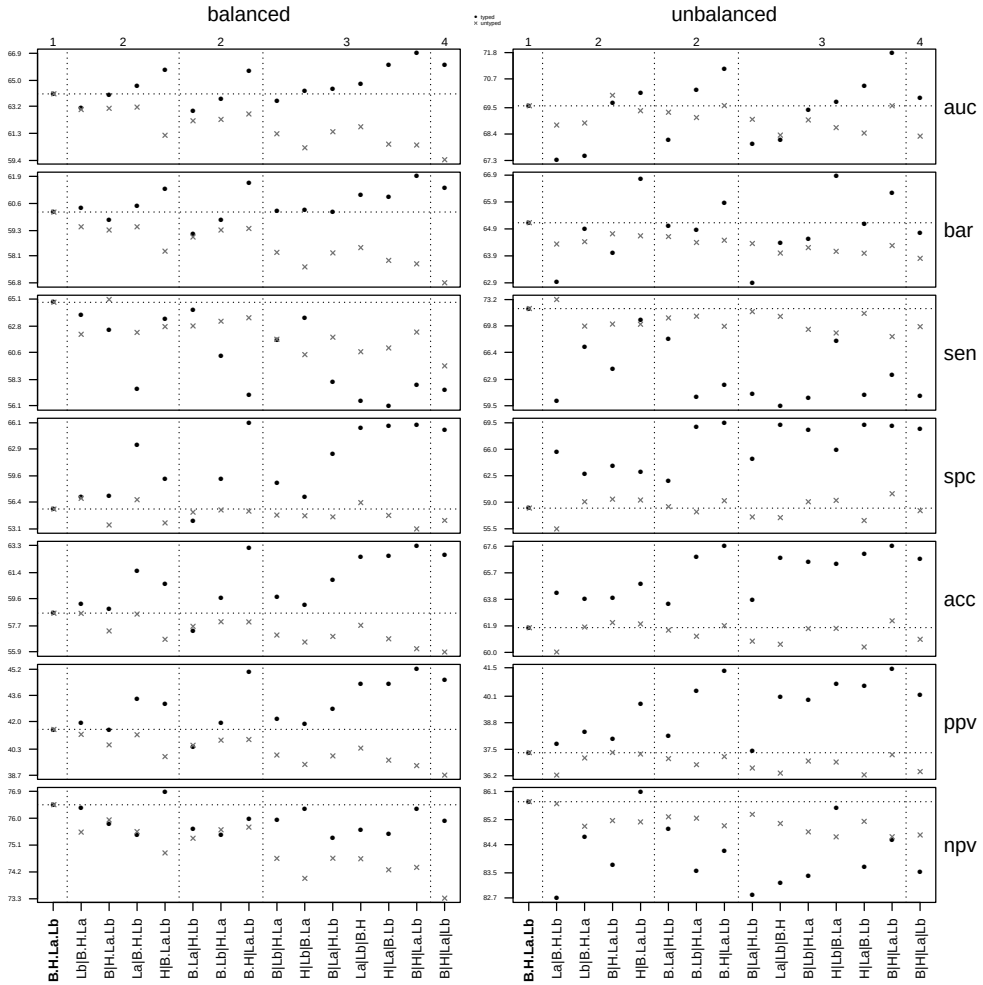


Figure 4.5: Balanced and unbalanced overall average performance overview. Performance overview of overall performance corresponding to the 15 distinct partitions w.r.t. the elementary subtype set $S_e = \{L_a, L_b, H, B\}$, that represents the subtypes lumA, lumB, HER2 and basal, respectively (Figure 4.2). The left panel corresponds to experiments involving the balanced compendia B, while the right panel corresponds to experiments involving the full unbalanced compendium D. In each panel the top numbers $\{1, 2, 3, 4\}$ indicate the number of different parts in each of the partitions, while the bottom line identifies the precise makeup of the various partitions e.g. the notation B|H|La.Lb indicates a partition into three parts, involving separate basal and HER2 groups, while having a combined luminal group. In each panel the coarsest partition is situated at the outer left, which corresponds to the baseline predictor (indicated in bold), that is, a single predictor that targets all samples. The most refined partition is situated at the outer right, which uses a separate predictor for each elementary subtype. A horizontal dotted line indicates the performance of the baseline predictors. Vertical dotted lines are used to group the partitions by their number of parts, as indicated by the top numbers. Results represent averages over 100 repeats. Rows represent seven frequently used performance indicators: area under curve (*auc*), balanced accuracy (*bar*), sensitivity (*sen*), specificity (*spc*), accuracy (*acc*), positive predictive value (*ppv*) and negative predictive value (*npv*). Performance for typed predictors is indicated with a dot, performance for untyped predictors with a cross.

		lumA	lumB	basal	HER2	overall
tp	auc	***61.5***	***65.0***	***60.6***	***74.7***	***66.1***
	bar	***56.3***	***60.8***	***56.7***	***71.5***	***61.3***
	sen	37.5	***71.7***	44.6	***75.9***	57.4
	spc	***75.1***	*49.8*	***68.8***	***67.2***	***65.2***
	acc	***62.8***	***57.0***	***60.9***	***70.0***	***62.7***
	ppv	***42.4***	***41.2***	***40.9***	***52.9***	***44.5***
	npv	*71.2*	***78.4***	*72.0*	***85.1***	***75.9***
up	auc	55.3	60.6	57.1	65.9	59.4
	bar	53.8	57.0	54.7	61.7	56.8
	sen	***56.3***	66.4	*48.1*	67.0	***59.5***
	spc	51.3	47.5	61.3	56.5	54.1
	acc	52.9	53.7	57.0	59.9	55.9
	ppv	36.1	38.1	37.9	43.1	38.7
	npv	70.7	74.7	70.9	77.9	73.3

Table 4.2: Subtype-specific performance overview (balanced compendia). Performance overview per elementary subtype: typed (tp) versus untyped (un) predictors on balanced compendia B. The highest value for a paired typed and untyped performance measure is set in italic. If the difference is significant (two sided paired t-test, $\alpha = 0.01$) the entry is set in bold. Values in the column *overall* correspond to the overall performance depicted in the left panel of Figure 4.5.

the partition. Similarly, untyped overall performance is achieved by employing four downsized versions of the baseline predictor, in which each predictor is constructed on an equal number of good and poor prognosis samples as their typed counterparts.

This is indeed one of the best performing partitions, with an associated overall *auc* and *bar* of 66.1% and 61.3% for the typed predictors, respectively, compared to 59.4% and 56.8% for the untyped predictors.

A more detailed overview corresponding to this partitioning with a breakdown of performance per subtype is given in Table 4.2. The subtype distribution of the training data indeed has a considerable impact on the performance of a predictor. Especially the HER2 group benefits from using a typed prediction rule with an *auc* and *bar* of 74.7% and 71.5%, respectively, for the typed predictors, compared to 65.9% and 61.7% for the untyped predictors. Results show an improvement for almost all other performance indicators as well when using typed predictors over untyped predictors, although for some subtypes untyped predictors achieve a higher sensitivity.

The best overall performance is obtained by typed prediction using a partition which has separate HER2 and basal groups, and a combined luminal group (Figure 4.5, left panel, second partition from the right). This partition gives an overall *auc* and *bar* of 66.9% and 61.9%, respectively, compared to 60.5% and 57.7% for the untyped predictors.

Note that coarser partitions involve predictors for compound subtypes that are constructed on larger sample sets compared to those in more refined partitions.

		lumA	lumB	basal	HER2	overall
tp	auc	***64.8***	***71.9***	***62.2***	***74.7***	***69.9***
	bar	***56.3***	***64.7***	***58.0***	***71.5***	***64.8***
	sen	***31.3***	74.6	50.0	*75.9*	60.8
	spc	81.3	***54.7***	***66.1***	***67.2***	***68.8***
	acc	74.6	***60.7***	***60.2***	***70.0***	***66.7***
	ppv	20.5	***41.8***	***45.6***	***52.9***	***40.2***
	npv	***88.5***	83.2	***69.9***	***85.1***	83.5
up	auc	63.0	70.2	50.4	60.3	68.3
	bar	54.6	62.3	50.9	57.5	63.8
	sen	19.9	***82.7***	***81.7***	74.9	***69.7***
	spc	***89.2***	41.9	20.1	40.2	57.9
	acc	***80.0***	54.3	42.4	51.5	60.9
	ppv	***22.4***	38.3	36.8	37.9	36.4
	npv	87.9	***84.8***	65.6	76.7	***84.7***

Table 4.3: Subtype-specific performance overview (unbalanced compendium). Performance overview per elementary subtype: typed (tp) versus untyped (un) predictors on the unbalanced compendium D. The highest value for a paired typed and untyped performance measure is set in italic. If the difference is significant (two sided paired t-test, $\alpha = 0.01$) the entry is set in bold. Values in the column *overall* correspond to the overall performance depicted in the right panel of Figure 4.5.

Increase in sample size can indeed be beneficial, as the baseline predictor, which is constructed on the largest training set possible under the given cross-validation scheme, is associated with the highest overall performance over all untyped predictors with an *auc* and *bar* of 64.1% and 60.2%, respectively (Figure 4.5). However, its performance is still lower than that obtained by using more refined typed prediction schemes. This clearly illustrates that a predictor trained on more samples without control for subtype distribution is not necessarily the optimal choice.
Finally, the increase in overall performance of typed predictors, as measured by *auc* and *bar*, is often accompanied by trading sensitivity for specificity. Compared to untyped predictors, typed predictors are generally associated with much higher specificity, yet lower sensitivity. Note that the highest sensitivity is in fact obtained by the baseline predictor.

Performance on unbalanced compendium

The right panel of Figure 4.5 reveals a similar pattern for typed and untyped prediction on an unbalanced compendium as seen in the left panel. Note that in contrast to the balanced sets B, the set D is unbalanced w.r.t. subtype distribution and is dominated by luminal samples (Table 4.1), hence performance on these samples drives overall performance. As expected, since most parts in the various partitions now contain a considerably larger number of samples compared to the balanced scenario, overall performance in terms of *auc* and *bar* improves. Similar to the balanced case, the highest overall performance

	lumA	lumB	basal	HER2	overall
auc	***68.6***	***72.7***	50.4	60.6	69.6
bar	51.8	63.2	49.5	58.1	***65.1***
sen	5.8	***87.8***	***86.8***	***84.9***	***72.0***
spc	***97.9***	38.6	12.2	31.3	58.2
acc	***85.6***	53.5	39.3	48.8	61.8
ppv	***29.9***	38.4	36.1	37.5	37.3
npv	87.1	***87.9***	62.1	80.9	***85.8***

Table 4.4: Baseline predictor performance. Baseline predictor performance on the unbalanced compendium D. Values are compared with those for the typed predictors in Table 4.3 and set in italic when higher. If the difference is significant (two sided paired t-test, $\alpha = 0.01$) the entry is set in bold.

is obtained by using a partition which has separate HER2 and basal groups, while using a combined luminal group. This partition has an *auc* and *bar* of 71.8% and 66.3%, respectively, which again outperforms the baseline predictor, which has an associated *auc* of 69.6 and 65.1%.

Table 4.3 is the unbalanced counterpart of Table 4.2. For the typed predictors an increase in sample size is indeed beneficial, as the *auc* and *bar* for all subtypes but HER2 increase. Note that the HER2 group in both the balanced and unbalanced case has the same size, hence its performance in the typed case remains unchanged. Furthermore, the most refined typed prediction scheme again outperforms its untyped counterpart, with an overall *auc* and *bar* of 69.9% and 64.8%, compared to 68.3% and 63.8%.

For the untyped predictors, however, the story is more complex. Table 4.3 shows a substantial gain in overall performance for the untyped predictors, compared to the untyped overall performance of Table 4.2, with an *auc* and *bar* of 68.3% and 63.8%, respectively, compared to 59.4% and 56.8% on the balanced compendia. Although we see a substantial improvement in *auc* for lumA and lumB, for basal and HER2 we observe a considerable deterioration. However, since luminal samples dominate the subtype distribution in the unbalanced case, overall performance for untyped prediction still improves quite strongly compared to the balanced scenario. In addition, a striking difference between the sensitivity and specificity of the lumA subtype compared to the other subtypes can be observed.

4.4.2 A dissection of the baseline performance

Table 4.4 presents a more detailed overview of how the baseline predictor obtains its performance. The baseline predictor shows an even more extreme difference between sensitivity and specificity, with a very high specificity for the lumA subtype of 97.9%, yet with a very low sensitivity of 5.8%. However, the sensitivity over the remaining subtypes is very high with values of 87.8%, 86.8%

and 84.9% for the subtypes lumB, basal, and HER2, respectively. Apparently, the unbalanced untyped predictors are biased to predict a good prognosis for lumA samples, yielding a very high specificity but very poor sensitivity for that subtype, and to predict a poor prognosis for the other subtypes, yielding a high sensitivity but a rather low specificity for them. Finally, we note the peculiar behavior of the *bar* performance indicator in an unbalanced setting. The overall *bar* is 65.1%, however, for *every* individual subtype the corresponding *bar* is less, even though they form a partition of the complete sample set D. The same phenomenon can be seen for the untyped predictors of Table 4.3.

4.5 Discussion

In van't Veer and Bernards (2008) it is suggested that the intrinsic breast cancer subtypes do not contain additional information for determining a patient's prognosis. They furthermore state that their value has been surpassed by that of prognostic gene-expression signatures such as the 70-gene signature, however, without quantifying these claims. This chapter presented a framework for building and quantifying the performance of typed and untyped predictors, inspired by the protocol proposed by Wessels *et al.* (2005). Our results show that the subtype distribution of the training data has a considerable impact on the behavior of a predictor and we provide strong evidence that event prediction can be improved by exploiting subtype information. The highest performance is obtained by partitioning the samples into separate basal and HER2 groups, while using a combined luminal group.

These results are in line with improved predictive power that was also reported using an intrinsic gene list (IGL) approach by Parker *et al.* (2009), which suggests a standardized gene set (PAM50) for subtype identification and event prediction. However, they only compare their subtype predictor with models based on standard clinicopathological parameters, such as estrogen receptor status and tumour size, and not with an untyped gene expression based predictor. The module-driven approach of Desmedt *et al.* (2008) has also been used to combine subtype-specific predictors in a fuzzy way with promising results (Haibe-Kains *et al.*, 2010). Although comprehensive, the latter work does not address influential factors like unequal class distributions or differences in the number of samples per subtype and presents its case for a single model, using a single partitioning scheme.

The module-driven approach was selected over the more common intrinsic gene list approach of Perou *et al.* (2000) because of favorable stability properties, which are extensively addressed in Haibe-Kains (2009). We stress that even though the exact method used to generate subtype information is of interest,

it is *not* the primary concern of this chapter, as here we are mainly interested in how typed and untyped prediction can be properly compared given the various forms of imbalance.

Sample size

As previously observed, stratification by subtype is accompanied by a sharp decrease in the number of samples available for predictor construction. Pairing typed predictors with untyped predictors offers the possibility to separately evaluate the influence of sample size and subtype information on classification performance. Our protocol incorporates two alternate views on sample size. Typed partitioning schemes involve multiple predictors, each targeting a specific subset of the entire sample set. Each typed predictor is paired with an untyped predictor, the construction of which involves an identical number of samples as for the typed predictor but with a subtype distribution that has been randomized such that it reflects the subtype distribution of the compendium. The advantage of matching sample size is that if subtyping would have no added value, paired typed and untyped predictors are expected to yield similar performance. Another view is provided by the comparison of typed predictors with the untyped baseline predictor in terms of overall performance. Prior to partitioning, all training sets are equally large. Hence, both typed and baseline predictor schemes involve the same total number of samples. According to both views typed predictors consistently outperform their untyped counterparts.
The potential to increase classification performance for breast cancer event prediction by combining data sets was addressed by van Vliet *et al.* (2008) which identified sample size as an important factor. In addition, it was observed that the performance on ER negative samples was much lower than achieved on ER positive samples, which matches well with the fact that the former group is substantially smaller than the latter. However, our work shows that when sample size is carefully controlled, performance differences between subtypes persist and cannot be ascribed solely to differences in sample size. For instance, basal samples, which are predominantly ER negative, appear an intrinsically more difficult set of samples to classify than HER2 samples.

Class imbalance

We performed an analysis on a set of balanced and unbalanced compendia by which we show that typed predictors consistently outperform their untyped counterparts. Especially the balanced scenario shows the potential of typed predictors. In an unbalanced setting, however, it may be more challenging to exploit subtype information for various reasons. Typed schemes attempt to

increase overall performance by predictors that perform well for all distinct parts. Such a strategy is not necessarily optimal in an unbalanced setting, as a predictor can be associated with a poor performance over all parts separately, yet can still give a reasonable overall performance over the union of these parts. This phenomenon is intimately related to the negative-positive class ratio and is perhaps easiest explained via the balanced accuracy rate (*bar*). The *bar* is defined as the average of the sensitivity and specificity, that is, $\text{bar} = \frac{1}{2} \cdot (\text{sen} + \text{spc}) = \frac{1}{2} \cdot (\frac{TP}{P} + \frac{TN}{N})$, where P and N denote the number of positive and negative samples, respectively, and TP and TN denote the true-positive and true-negative assignments made by a predictor. The *bar* score can be highly sensitive to the negative-positive class ratio in a subtle way. This becomes clear when rewriting the *bar* as a weighted accuracy measure

$$\text{bar} = \frac{w_P \cdot TP + w_N \cdot TN}{w_P \cdot P + w_N \cdot N},$$

with weights $w_P = \frac{N}{P}$ for the positive instances and $w_N = 1$ for the negative instances. Depending on the negative-positive class ratio, an error on a positive case is weighted differently from an error on a negative case. Hence, given the different negative-positive class ratios for different subtypes and for the whole compendium (Table 4.1), the same errors are weighted differently in the unbalanced compendium. For instance, the negative class is strongly overrepresented in the lumA subtype. In terms of *bar* the misclassification of a positive example in this case is extremely costly, as expressed by a *bar* of merely 51.8% in Table 4.4. The overall *bar*, however, weighs its errors very differently which results in a more optimistic *bar* of 65.1%. The latter example indicates the importance of proper stratification when comparing performances between groups.

In conclusion, we have presented a novel experimental protocol that allows for a proper comparison between typed and untyped predictors. We performed a comprehensive analysis of our methodology on a large breast cancer compendium and presented an analysis for balanced and unbalanced scenarios, which clearly reveal the potential of typed prediction. In both scenarios the highest overall performance was obtained by a typed partition which had separate HER2 and basal groups, while using a combined luminal group. In the balanced scenario it was observed that certain subtypes appear intrinsically more challenging as performance rates differ between subtypes. In an unbalanced setting it can be more difficult to exploit subtype information as the performance of certain subtypes can dominate overall performance. In addition, in such a scenario comparisons between predictors can be obscured by differences in sample size or class distribution. In our protocol sample size,

class and subtype distributions are carefully controlled, which combined with the systematic pooling steps offers a rich view on the value of subtypes for event prediction.

CHAPTER 5

DECOMPOSITION OF PERFORMANCE MEASURES UNDER SUBTYPES

5.1 Abstract

For some classification tasks the corresponding data can be partitioned into disjoint subsets based on some attribute, for example a disease subtype. It then seems logical to train a classifier with the same classes as the original classification problem for each subtype separately, such that the performance per subtype is optimized. Unfortunately, the influence of the subtype performances on the aggregated overall performance depends strongly on the performance measure used and can be very counterintuitive. We show that for some performance measures (*e.g.* classification accuracy, precision, recall, F_1) the aggregated performance is a simple linear combination of subtype performances. In these cases, improving the performance of a subtype-specific classifier implies that the overall performance improves. However, for other performance measures (*e.g.*, balanced accuracy rate, area under the ROC curve) and also for performance measures in survival analysis (concordance index), additional cross terms appear in the aggregation of the subtype performances. These cross terms are heavily dependent on both the overall class imbalance and the subtype class imbalances. For these measures, improving subtype performances may actually result in a decrease of the overall performance. [1]

[1] This work was published as: DMJ Tax*, HMJ Sontrop*, MJT Reinders, PD Moerland (2014). The effect of aggregating subtype performances depends strongly on the performance measure used. Proceedings of International Conference on Pattern Recognition (ICPR) 2014. IEEE. *Contributed equally

5.2 Introduction

In general, for dichotomous classification problems a model is trained on all available data, except some left out validation data. In some situations though, each datapoint is associated with exactly one out of a heterogeneous set of subtypes. A typical example in a biomedical setting is breast cancer event prediction, where one attempts to predict whether a tumor will metastasize or not (van't Veer *et al.*, 2002). Here, each breast cancer sample can often be assigned to one out of four different tumor subtypes (Perou *et al.*, 2000). Another example is the problem of face detection in arbitrary images, where one may distinguish subtypes of faces according to gender, presence of glasses, etc. In these cases, instead of using a single monolithic classifier to predict a binary class label, better dichotomous classifiers might be constructed for each suptype separately, targeting what are thought to represent more homogeneous sets of samples (Haibe-Kains *et al.*, 2010; Sontrop *et al.*, 2011). The question we address in this chapter is how aggregating performances per subtype affects the overall performance, in particular when subtypes show an imbalance with respect to the class label distribution that is very different from the overall class imbalance.

Class imbalance is a well known problem in classification, especially for detection problems, where a rare class has to be discriminated from a large background class (Japkowicz and Stephen, 2002). An extreme case occurs in retrieval applications, where just a few positive items have to be retrieved from millions of background items. A performance comparison using a measure like the classification accuracy has the disadvantage that a classifier that assigns all objects to the large background class is associated with a high performance. To avoid this problem, more appropriate performance measures have been designed, like the balanced accuracy rate (BAR), the area under the ROC curve (AUC) or the F-measure (Wessels *et al.*, 2005; Bradley, 1997; van Rijsbergen, 1979).

When the data for a classification problem can be partitioned into subtypes, the subtypes typically have different class imbalances. It therefore seems natural to use performance measures that are insensitive to class imbalance in this case as well. However, in this chapter we show that even when a performance measure is used that is supposed to be insensitive to class imbalance, the combined overall performance might depend on the subtype class imbalances. Even worse, for some measures the overall performance can be arbitrarily good or bad, depending on the class imbalances of the subtypes.

5.3 Aggregation of performance measures

We assume that for a given two-class classification problem, each datapoint is associated with exactly one out of a set of s subtypes. The dataset D can then be partitioned into disjoint subsets S_i, each having P_i positive and N_i negative objects for subtype i. The total number of positive and negative objects in D is $P = \sum_{i=1}^{s} P_i$, and $N = \sum_{i=1}^{s} N_i$, respectively. For each of the subsets S_i, a separate subtype-specific binary classifier f_i is fitted for the same two classes. Of the positive objects of subset S_i, TP_i are classified as positive, and FN_i are classified as negative. Similar definitions hold for the negative objects of subset S_i. In Table 5.1 the relevant notation and definitions of performance measures are given. For each definition its subtype-specific equivalent is indexed with a subscript i, for example, acc_i denotes the accuracy on subtype i.

5.3.1 The simple case: linear combination

The aggregation of subtype performances π_i is straightforward when the performance measure can be written as the ratio of two positive-valued terms A and B, and when both terms can be decomposed as the sum over their subtype-specific counterparts: $A = \sum_{i=1}^{s} A_i$ and $B = \sum_{i=1}^{s} B_i$, respectively. Of the performance measures in Table 5.1, the accuracy, precision, recall, specificity, negative predictive value and the F-measure are of this type. Then:

$$\pi = \frac{A}{B} = \sum_{i=1}^{s} \frac{A_i}{B} = \sum_{i=1}^{s} \frac{B_i}{B}\frac{A_i}{B_i} = \sum_{i=1}^{s} \frac{B_i}{B}\pi_i. \tag{5.1}$$

The overall performance π is rewritten in terms of the subtype-specific performances π_i, weighted by the fraction $\frac{B_i}{B}$. Since this weight is normalized (between 0 and 1, and summing up to 1) the overall performance π is somewhere between the lowest and highest subtype performance π_i. Another consequence is that improving the performance of one of the subtype-specific classifiers (while keeping the performance of the others equal) will pay off, in the sense that the aggregated performance will also increase.
Other performance measures, like the balanced accuracy rate, the AUC and the concordance index, are not of this form. In the following subsections we derive the formulae for these three measures.

5.3.2 Balanced accuracy rate

To aggregate the overall balanced accuracy rate (BAR) from the subtype balanced accuracy rates BAR_i, we start by applying (5.1) to the definition of

P,N	nr. of positive, negative objects	
TP,TN	nr. of correctly classified positives, negatives	$TP = \sum_i^s TP_i,\ TN = \sum_i^s TN_i$
FP,FN	nr. of incorrectly classified positives, negatives	$FP = \sum_i^s FP_i,\ FN = \sum_i^s FN_i$
$\mathbf{x}^+$, $\mathbf{x}^-$	positive, negative object	
$f(\mathbf{x})$	output of classifier f for object $\mathbf{x}$	
t_i	survival time for object $\mathbf{x}_i$	
Ω	set of all pairs $(\mathbf{x}_k, \mathbf{x}_l)$ for which $t_k < t_l$	
acc	accuracy	$\frac{TP+TN}{P+N}$
prec	precision (or positive predictive value)	$\frac{TP}{TP+FP}$
rec	recall (or sensitivity)	$\frac{TP}{P}$
spec	specificity	$\frac{TN}{N}$
npv	negative predictive value	$\frac{TN}{TN+FN}$
F_1	F-measure	$2\frac{\text{prec}\cdot\text{rec}}{\text{prec}+\text{rec}} = 2\frac{TP}{2TP+FN+FP}$
BAR	balanced accuracy rate	$\frac{1}{2}\frac{TP}{P} + \frac{1}{2}\frac{TN}{N}$
AUC	area under the ROC curve	$\Pr[f(\mathbf{x}^+) > f(\mathbf{x}^-)]$
c	concordance index	see Eq. 5.11

Table 5.1: Notation and definitions used in this chapter.

the BAR:

$$\begin{aligned} \text{BAR} &= \frac{1}{2}\frac{TP}{P} + \frac{1}{2}\frac{TN}{N} && (5.2)\\ &= \frac{1}{2}\sum_{i=1}^{s}\frac{TP_i}{P} + \frac{1}{2}\sum_{i=1}^{s}\frac{TN_i}{N} && (5.3)\\ &= \frac{1}{2}\sum_{i=1}^{s}\frac{P_i}{P}\frac{TP_i}{P_i} + \frac{1}{2}\sum_{i=1}^{s}\frac{N_i}{N}\frac{TN_i}{N_i}. && (5.4) \end{aligned}$$

By adding and subtracting the term $\frac{TN_i}{N_i}$ from the first term in (5.4), and after rearrangement we obtain:

$$\begin{aligned} \text{BAR} &= \frac{1}{2}\sum_{i=1}^{s}\frac{P_i}{P}\left(\frac{TP_i}{P_i}+\frac{TN_i}{N_i}-\frac{TN_i}{N_i}\right) \\ &\quad +\frac{1}{2}\sum_{i=1}^{s}\frac{N_i}{N}\frac{TN_i}{N_i} \\ &= \sum_{i=1}^{s}\frac{P_i}{P}\text{BAR}_i+\frac{1}{2}\sum_{i=1}^{s}\left(\frac{N_i}{N}-\frac{P_i}{P}\right)\frac{TN_i}{N_i}. \end{aligned} \tag{5.5}$$

Thus, the overall BAR is a linear combination of the BAR_i of the subtypes, with weights that depend on the fraction of positive objects in subtype S_i, corrected with a term that depends on both the subtype imbalances as on the overall class imbalance.
If the correction term $\left(\frac{N_i}{N}-\frac{P_i}{P}\right)\frac{TN_i}{N_i}$ is positive (negative), the overall performance can be larger (smaller) than any of the individual subtype performances. This becomes more pronounced when the relative imbalance between the classes, i.e. the difference between N_i/N and P_i/P, becomes larger. In particular when a class is strongly overrepresented in one subtype, but underrepresented in another subtype, the aggregated balanced accuracy rate can become completely different from the BAR_i of the subtypes. Therefore, improving the BAR for one subtype might result in worse overall performance even when keeping the performance of the other subtype-specific classifiers equal. Only when the negative class is strongly overrepresented in subtype i (or equivalently, the positive class is strongly underrepresented in subtype i) and TN_i/N is high, the overall performance improves when the BAR_i improves.
Note that we could also have expanded the second term in equation (5.4) by adding and subtracting $\frac{TP_i}{P_i}$. This would have resulted in:

$$\text{BAR} = \sum_{i=1}^{s}\frac{N_i}{N}\text{BAR}_i+\frac{1}{2}\sum_{i=1}^{s}\left(\frac{P_i}{P}-\frac{N_i}{N}\right)\frac{TP_i}{P_i}. \tag{5.6}$$

By taking the average of decompositions (5.5) and (5.6) we obtain:

$$\begin{aligned} \text{BAR} &= \frac{1}{2}\sum_{i=1}^{s}\left(\frac{P_i}{P}+\frac{N_i}{N}\right)\text{BAR}_i \\ &\quad +\frac{1}{4}\sum_{i=1}^{s}\left(\frac{P_i}{P}-\frac{N_i}{N}\right)\left(\frac{TP_i}{P_i}-\frac{TN_i}{N_i}\right). \end{aligned} \tag{5.7}$$

This expression makes clear that when a subtype-specific classifier has much higher performance on the positive class than on the negative class, i.e. $\left(\frac{TP_i}{P_i} - \frac{TN_i}{N_i}\right)$ is large, and this subtype covers a relatively larger fraction of the positive objects than of the negative objects, i.e. $\left(\frac{P_i}{P} - \frac{N_i}{N}\right)$ is large as well, then the overall performance will be even better than is expected based on the subtype performances BAR_i only.

5.3.3 Area under the ROC curve

The area under the ROC curve evaluates the ranking of the objects in a dataset and can be defined as $\text{AUC} = \Pr[f(\mathbf{x}^+) > f(\mathbf{x}^-)]$, i.e. the probability that, when a random object $\mathbf{x}^+$ is drawn from the positive class and a random object $\mathbf{x}^-$ from the negative class, the positive object has a higher classifier output $f(\mathbf{x}^+)$ than the negative object $f(\mathbf{x}^-)$ (Bradley, 1997). This probability can be estimated from a sample with positive objects $\mathbf{x}_k^+$ and negative objects $\mathbf{x}_l^-$ using the estimator:

$$\text{AUC} = \frac{1}{PN} \sum_{k=1}^{P} \sum_{l=1}^{N} \mathbb{I}(f(\mathbf{x}_k^+) > f(\mathbf{x}_l^-)), \tag{5.8}$$

where $\mathbb{I}(a > b)$ is the indicator function that equals 1 when the condition $a > b$ is satisfied, and 0 otherwise. Since the AUC is rank-based, the offset and scale of the outputs $f(\mathbf{x})$ do not matter in the computation of the AUC of a single classifier. However, in our setup where the outputs of different subtype-specific classifiers f_i are combined, scaling and offset values become important.
The overall AUC (5.8) can be decomposed as follows for $\mathbf{x}_k^+ \in S_i$ and $\mathbf{x}_l^- \in S_j$:

$$\begin{aligned} \text{AUC} = & \sum_{i=1}^{s} \sum_{j=1}^{s} \frac{P_i N_j}{PN} \frac{1}{P_i N_j} \times \\ & \sum_{k=1}^{P_i} \sum_{l=1}^{N_j} \mathbb{I}[f_i(\mathbf{x}_k^+) > f_j(\mathbf{x}_l^-)] \qquad (5.9) \\ = & \sum_{i=1}^{s} \frac{P_i N_i}{PN} \text{AUC}_i + \sum_{(i,j):i \neq j}^{s} \frac{P_i N_j}{PN} \text{AUC}_{ij}, \qquad (5.10) \end{aligned}$$

where $\text{AUC}_{ij} = \Pr[f_i(\mathbf{x}_k^+) > f_j(\mathbf{x}_l^-)]$ and $\text{AUC}_i = \text{AUC}_{ii}$. Thus, similar to the BAR, the overall AUC is a weighted combination of the subtype AUCs, but extended with additional cross terms. These cross terms depend on the

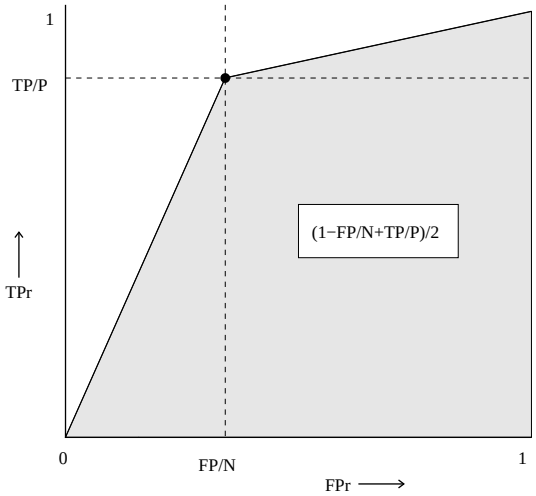


Figure 5.1: Approximation of the ROC curve when a single operating point on the curve is given, as defined by the false positive rate FP/N and the true positive rate TP/P.

rankings of the outputs of pairs of subtype-specific classifiers for the positive and negative objects. When the subtype-specific classifiers are not properly calibrated, these rankings can become arbitrarily good or bad. It is even possible to construct examples where the individual AUCs become 1, but where the total AUC still becomes almost 0.5. This is an example of Simpson's paradox where overall trends in the data are reversed in subsets of the same data (Moore, 1996).

It is hard to make general statements about AUC_{ij}, since it depends on the classifier output distributions for both subtypes i and j. As an approximation, let us assume that the outputs of the subtype-specific classifiers f_i are normalized, in the sense that their decision boundary is at a fixed position, say $f_i(\mathbf{x}) = 0$ for $\mathbf{x}$ on the decision boundary. Objects with output larger than 0 are assigned to the positive class, and otherwise to the negative class. The TP_i and FP_i of the subtype-specific classifiers therefore define one point on the ROC curve. The remainder of the ROC curve can be approximated by straight lines to $(0, 0)$ and $(1, 1)$, as shown in Figure 5.1, see also Cortes and Mohri (2004). This approximation is often on the pessimistic side, since it is based on the assumption that the erroneous objects are distributed evenly among the correct objects. In most situations erroneous objects are closer to the decision boundary, resulting in a more concave curve and a bit higher value for the AUC.

Under the approximation with straight lines, the area under the ROC curve can be written as:

$$\text{AUC} = \frac{1}{2}\left(1 - \frac{FP}{N} + \frac{TP}{P}\right) = \frac{1}{2}\left(\frac{TN}{N} + \frac{TP}{P}\right)$$

The decomposition of the AUC with the straight-line approximation, therefore, results in an identical result as that of the balanced accuracy rate in equations (5.5)-(5.7) and is subject to the same problem.

5.3.4 Concordance index

In survival data, each object $\mathbf{x}_i$ has a corresponding real-valued survival time t_i instead of a discrete label. To measure the performance of a predictor f, the concordance index tests how often the ranking as obtained by the true survival times corresponds to the ranking obtained by f. Define Ω as the set of all pairs $(\mathbf{x}_k, \mathbf{x}_l)$ for which $t_k < t_l$. The concordance index is a generalization of the AUC and is defined as:

$$c = \frac{1}{|\Omega|} \sum_{(k,l)\in\Omega} \mathbb{I}(f(\mathbf{x}_k) < f(\mathbf{x}_l)). \tag{5.11}$$

The aggregated concordance index can be decomposed as follows:

$$\begin{aligned} c &= \sum_{i=1}^{s}\sum_{j=1}^{s} \frac{|\Omega_{ij}|}{|\Omega|} \frac{1}{|\Omega_{ij}|} \sum_{(k,l)\in\Omega_{ij}} \mathbb{I}(f_i(\mathbf{x}_k) < f_j(\mathbf{x}_l)) \\ &= \sum_{i=1}^{s} \frac{|\Omega_i|}{|\Omega|} c_i + \sum_{(i,j):i\neq j}^{s} \frac{|\Omega_{ij}|}{|\Omega|} c_{ij}. \end{aligned} \tag{5.12}$$

where the set Ω is partitioned into disjoint subsets Ω_{ij} for subtypes i and j, such that Ω_{ij} contains all pairs $(\mathbf{x}_k \in S_i, \mathbf{x}_l \in S_j)$ for which $t_k < t_l$ and where c_{ij} is the corresponding concordance index. Furthermore, we define the special cases $\Omega_i = \Omega_{ii}$ and $c_i = c_{ii}$. Similar to the AUC (5.10), the overall concordance index is a weighted sum of the subtype concordances, extended with cross terms c_{ij}. The overall concordance index therefore shows the same characteristics as the overall AUC, in that the cross terms can influence the overall concordance index in both a positive and a negative way.

	Subtype	P_i, N_i	$TP_i/P_i, TN_i/N_i$	BAR_i	Overall BAR
I	1	99, 1	1.0, 0.0	0.5	
	2	1, 99	0.0, 1.0	0.5	0.99
IIA	1	200, 10	0.9, 0.5	0.7	
	2	50, 50	0.9, 0.5	0.7	0.7
IIB	1	200, 10	0.6, 1.0	0.8	
	2	50, 50	0.9, 0.5	0.7	0.62
IIC	1	200, 10	0.9, 0.5	0.7	
	2	50, 50	0.5, 0.9	0.7	0.83

Table 5.2: Balanced accuracy rates on an artificial example.

5.4 Experimental results

Here we illustrate the counterintuitive results when aggregating subtype BARs and AUCs using a number of toy examples and a real-world dataset on breast cancer prognosis.

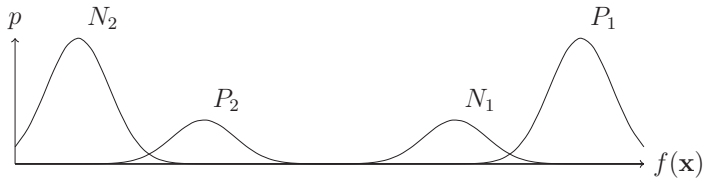
5.4.1 Toy examples: balanced accuracy rate

In Table 5.2 the BAR performances on two artificial examples are shown. In the first example, the two classes in both subtypes are heavily imbalanced ($P_1 = 99, N_1 = 1$ for the first subtype, and $P_2 = 1, N_2 = 99$ for the second subtype). The classifier for the first subtype assigns all objects to class 1. This results in a very good performance on class 1 (obtaining a true positive rate of 1), but a very poor one on class 2 (with a true negative rate of 0). For the second subtype, the reverse happens, and all objects are assigned to the negative class. The BAR for both subtypes is therefore 0.5, suggesting very poor overall performance on this problem. The picture changes when the overall BAR is computed: the class imbalance vanishes, and the overall BAR becomes 0.99. This example clearly shows that the subtype performances may not be indicative for the overall performance at all, even if a performance measure insensitive to class imbalance is used.

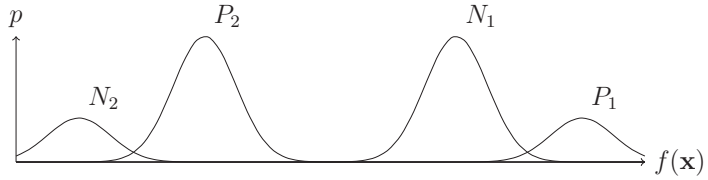
In the second example, the two classes in the first subtype are heavily imbalanced ($P_1 = 200, N_1 = 10$), while in the second subtype the classes are balanced ($P_2 = N_2 = 50$). The two rows labeled IIA in Table 5.2 define the baseline performance. For both subtypes the true positive rate and the true negative rate are 0.9 and 0.5, respectively. This results in a subtype-specific BAR of 0.7 in both cases. The overall BAR also equals 0.7, since in this case $\frac{TP_i}{P_i} = \frac{TN_i}{N_i}$ and the last term in (5.7) disappears. In the next two rows (IIB)

the BAR_1 on the first subtype is improved from 0.7 to 0.8. Unfortunately, due to the large class imbalance and the abundance of positive objects in subtype 1, the overall BAR decreases to 0.62. This illustrates that improving the subtype performance can deteriorate the overall performance. In the last two rows (IIC) the subtype-specific BAR performances are kept constant, but the classifier in subtype 2 is changed in such a way that the per class accuracies are swapped. The overall BAR now improves significantly to 0.83. Taken together, these examples show that by changing the subtype performances in subtypes with large class imbalance, the overall BAR can either improve or deteriorate.

5.4.2 Toy examples: area under the curve



(a) The class imbalance in the subtypes results in a high overall AUC.



(b) The class imbalances in the subtypes are counterproductive and lower the overall AUC (i.e. Simpson's paradox).

Figure 5.2: Distribution of the combined outputs of two subtype classifiers. Outputs of subtype 1 (i.e., P_1 and N_1) are always much higher than those of subtype 2 (i.e., P_2 and N_2).

Assume that we have two subtypes, and the classifier outputs are very poorly calibrated: all outputs for subtype 1 are much larger than for subtype 2. Then $\text{AUC}_{1,2} = \Pr[f_1(\mathbf{x}_1^+) > f_2(\mathbf{x}_2^-)] = 1$, $\text{AUC}_{2,1} = \Pr[f_2(\mathbf{x}_2^+) > f_1(\mathbf{x}_1^-)] = 0$ and the overall AUC of equation (5.10) simplifies to $\text{AUC} = \sum_{i=1}^{2} \frac{P_i N_i}{PN} \text{AUC}_i + \frac{P_1 N_2}{PN}$. For this case, Figure 5.2(a) shows that when the positive class is much larger than the negative class in subtype 1, and vice versa in subtype 2, then $P_1 N_2$ is large, and the AUC improves a lot. In this situation, the combined AUC

	BAR	AUC
lumA	56.30 (1.98)	64.76 (1.65)
lumB	64.66 (1.66)	71.91 (1.34)
basal	58.03 (2.06)	62.21 (1.66)
HER2	71.52 (1.23)	74.66 (0.82)
overall	64.75 (0.93)	69.91 (0.63)

Table 5.3: The BAR and AUC for the four breast cancer subtypes, and the overall BAR and AUC (in percentages). Values between brackets indicate the standard deviation over 10-fold crossvalidation.

	lumA	lumB	basal	HER2
P_i/P	0.183	0.411	0.249	0.157
$N_i/N - P_i/P$	0.228	-0.085	-0.098	-0.046

Table 5.4: Terms determining the influence of the subtype BARs on the overall BAR.

can be much higher than the individual AUCs. On the other hand, when the positive class in subtype 1 and the negative class in subtype 2 are very small, the total AUC may become low (Figure 5.2(b)).

5.4.3 Subtype-specific breast cancer event prediction

Breast cancer event prediction is a challenging classification problem in which one attempts to predict whether a breast tumor will metastasize or not, given a patient's gene expression data as measured on microarrays (van't Veer *et al.*, 2002). It appears that at the molecular level breast cancer consists of a heterogeneous set of subtypes, each having a potentially different prognosis. In Perou *et al.* (2000) four types were distinguished: luminal A (lumA), luminal B (lumB), basal, and HER2. Using a model-based clustering scheme to identify subtypes, we recently constructed subtype-specific nearest centroid classifiers on a large breast cancer gene expression dataset (Sontrop *et al.*, 2011).

In Table 5.3 the balanced accuracy rate and the area under the ROC curve performances that we obtained are shown. The first four rows give the subtype performances, the fifth row indicates the overall performance where all the outcomes are combined to get the overall BAR or AUC. Equation (5.5) shows that the terms P_i/P and $N_i/N - P_i/P$ determine the influence of the subtype BAR_i and the subtype true positive rates, respectively, on the overall BAR value. In Table 5.4 these terms are shown for the four subtypes (averaged over 10 folds).

For subtype lumA the second term becomes positive, indicating that an increase in true negative rate for this subtype-specific classifier might improve the overall BAR and AUC. By adding a suitably chosen offset to the output of the lumA classifier (and therefore artificially changing the decision threshold), the true negative rate is increased from 0.726 to 0.989. Indeed the BAR for subtype lumA decreases from 56.3 to 51.93, while the BAR for the other subtypes remains equal and the overall BAR improves from 64.75 to 65.95. Note that by adding an offset to the output of the lumA classifier, the $\mathrm{AUC}_{\mathrm{lumA}}$ does not change. But when the shifted lumA classifier output is combined with the outputs of the other subtype-specific classifiers, the cross terms from (5.10) appear and the overall AUC improves from 69.91 to 71.72.

5.5 Discussion and conclusions

In this chapter we have shown that when for a classification problem the data is partitioned into subtypes for which subtype-specific classifiers are fitted, the aggregated overall performance can deviate much from the individual subtype performances. Depending on the type of performance measure, the class imbalance in the subtypes and the actual subtype performances, improving the performance on a subtype may even deteriorate the overall performance.

The classification accuracy, the F-measure, negative/positive predictive value, sensitivity and specificity, all show a reasonable behavior, in that the overall performance is a weighted linear combination of the individual performances. The overall performance will therefore lie between the worst and best subtype performance. Furthermore, improving the performance for a subtype will improve the overall performance. It therefore pays off to fit individual models that are as good as possible on their own.

For other performance measures, like the balanced accuracy rate, the AUC, or in the case of survival models, the concordance index, the overall performance is not just a linear combination of subtype performances. In this case the formulae for the overall performance contain additional cross terms. It can then happen that the overall performance becomes higher than the highest, or lower than the lowest subtype performance. Even worse, improving the performance for a subtype will not automatically improve the overall performance. A subtype with a large relative imbalance between the two classes can have a large influence on the overall performance, but this may not be clearly visible in the subtype BAR or AUC performance (in particular because the measures like balanced accuracy rate and AUC are used to deal with imbalanced classification

problems).

This should be taken into account when models are fitted and tested on data containing subtypes. Especially so, if one wants to also compare performances between subtype-specific classifiers. Take as an example the BAR, with the decomposition given in (5.7). The unpredictable element in comparing and combining subtype BARs is the second term in (5.7). In order to reduce the influence of the cross term two approaches can be taken. The first approach is to make $\left(\frac{P_i}{P} - \frac{N_i}{N}\right) = 0$, or equivalently $\frac{P_i}{N_i} = \frac{P}{N}$. This means that the class imbalances in the subtypes have to be identical to the overall class imbalance. In general, this means that a stratified sampling scheme has to be used to make sure that the subtype class imbalance is equal to $\frac{P}{N}$, similar to what is customary in the context of classifier performance evaluation using cross-validation (Kohavi, 1995). The second approach is to make $\left(\frac{TP_i}{P_i} = \frac{TN_i}{N_i}\right)$, thus making sure that the operating point of the subtype-specific classifiers is set to have equal error rates per class. In particular for heavily imbalanced classes this may require some extra effort. The standard way of doing this is to (*i*) resample the data such that the classes are balanced in order to avoid class imbalance, (*ii*) recalibrate the decision threshold in order to obtain an equal error for both classes. When one of the two above-mentioned approaches is used, it is possible to improve the overall performance by optimizing the subtype performances, resulting in good overall and subtype performances.

CHAPTER 6

DISCUSSION

6.1 Easy versus complex

In the protocols and methodologies used in this dissertation often multiple options were available to perform a specific step, e.g. feature selection, choice of classifier, batch correction technique etc. Simple techniques often worked as well as more complex alternatives. In some of our experiments complex approaches even led to far inferior results. Below we describe two insights from the additional experiments that we performed.

6.1.1 Multivariate feature selection by combinatorial optimization

Instead of relying on *univariate* methods as used in Chapters 2 en 4, we have done extensive experiments to investigate whether the performance of event prediction could be improved by *multivariate* feature selection. In these experiments, multivariate feature selection was driven by advanced local search metaheuristics such as *simulated annealing* (Aarts and Korst, 1990), adaptive memory programming techniques such as *tabu search* (Glover *et al.*, 1997) and *genetic algorithms* (Goldberg and Holland, 1988). In our experiments the prediction problem was formulated as a combinatorial optimization problem, i.e. for a given type of predictor, we look for the best feature set to use during prediction. Even though metaheuristic frameworks are extremely powerful in identifying high quality solutions for many hard combinatorial optimization problems (Aarts and Lenstra, 1997), it proved difficult to find practical objective functions of which the optimization did *not* lead to overfitting. Simple univariate techniques, such as forward filtering or backward feature elimination schemes (Wessels *et al.*, 2005), proved to be far superior, both in terms of

computational cost and generalization performance of the model. Until better optimization objectives can be formulated, approaches such as suggested by Schaffer *et al.* (2005), are ill-advised on this type of data. Instead of using combinatorial optimization, multivariate feature selection may be performed by simpler methods, e.g. *top-scoring pairs* (Geman *et al.*, 2004), or variants of these (Tan *et al.*, 2005). In our experiments such schemes, however, did not lead to improvements over simple univariate techniques (Chapter 2). Similar conclusions on univariate versus multivariate gene selection techniques for the classification of cancer datasets were reported by Lai *et al.* (2006).

6.1.2 Advanced data processing techniques

As noted by Leek *et al.* (2010), an often overlooked complication with high-throughput studies are *batch effects*. These occur because measurements are affected by laboratory conditions, reagent lots and differences in personnel. Especially when combining data from multiple experiments it becomes critical that potential batch effects are properly addressed. Unfortunately, the assignment of samples into batches is often non-trivial. Batches can be assigned based on differences in array designs, technologies and/or institutes. We refer to such batches as *processing group* based batches. These, however, only capture part of the unwanted technical variation. Besides differences due to changes in measurement technologies and/or institutes, microarray measurements are also influenced by subtle changes in experimental conditions. As it is difficult to control experimental conditions over large time spans, batch effects are also related to the processing times of an experiment. We refer to such batches as *processing time* based batches. As changes in experimental conditions can be subtle the experimenter may not even be aware of such changes. As a consequence such changes will not be annotated and the corresponding batch vector essentially becomes a latent variable.

Figure 6.1 *(facing page)*: Example of processing times and processing groups. Graphical overview of the partition of datasets considered in this dissertation into *processing group* and *processing time* based batches (Leek *et al.*, 2010). Columns correspond to processing groups, while rows represent distinct scan dates. Processing groups were based on descriptions found in the associated literature. Each processing groups was subsequently divided into processing time related batches, based on scan date information, similar to McCall and Irizarry (2011). Hybridizations were performed over a period of 7 years, i.e. from 2002-2009. Horizontal black solid lines divide the scan dates by year. Batches are colored by size, indicated by the legend at the right hand side. Note that not every array could be assigned to a processing time related batch (see red dots). Although the figure provides an impression of the time structure, only actual scan dates are shown, i.e. days on which no arrays were scanned were omitted. Hence, the fact that batches are displayed close together does not necessarily imply that the corresponding scan dates were close.

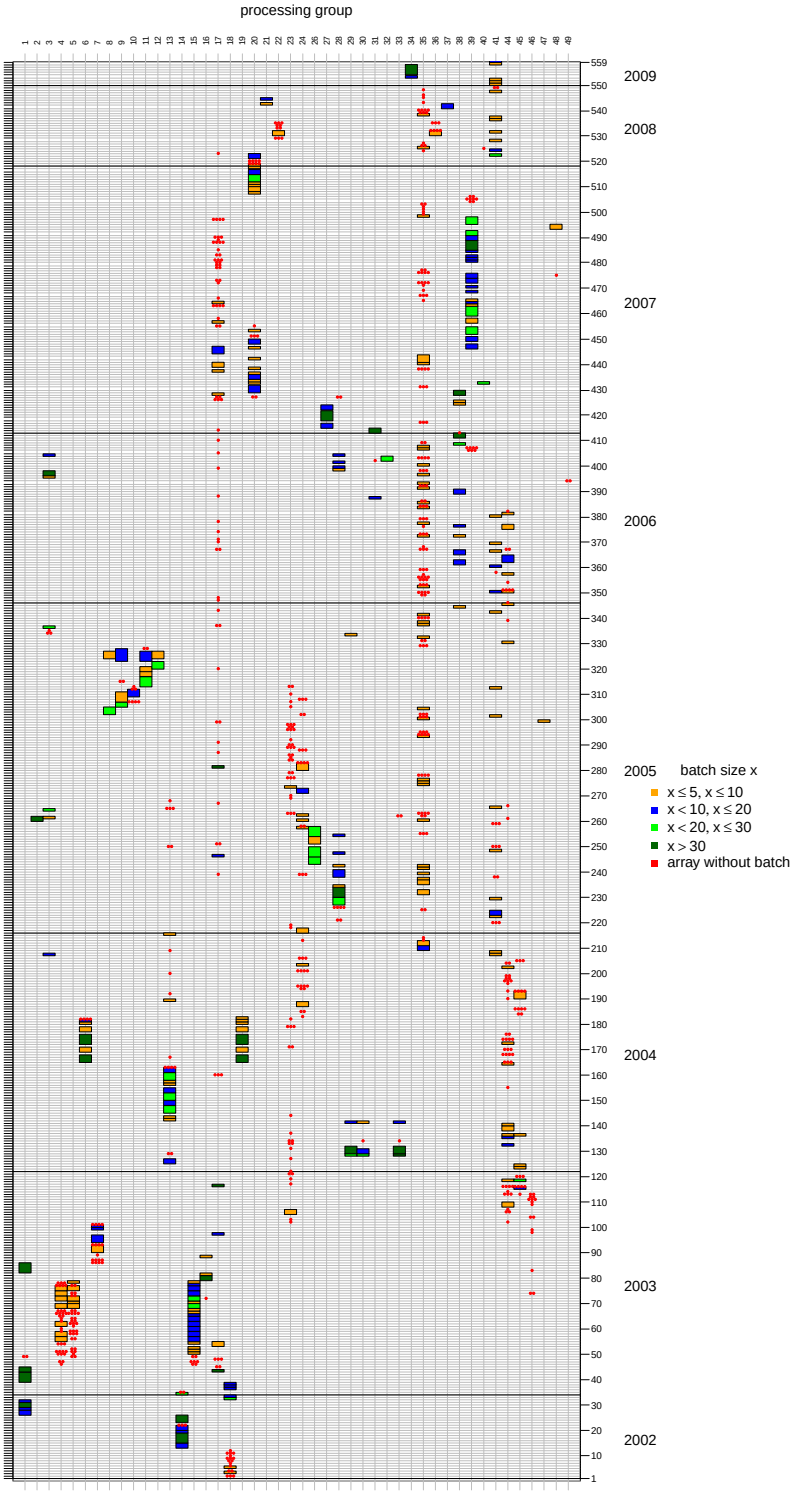

processing group
2009
2008
2007
2006
2005
2004
2003
2002
batch size x
x ≤ 5, x ≤ 10
x < 10, x ≤ 20
x < 20, x ≤ 30
x > 30
array without batch

Figure 6.1 on the previous page shows the result of a painstaking investigation of the literature on batch information on processing groups and processing times for a large number of cohorts considered in this dissertation. Processing time-based batches were based on the scan dates of the arrays, as suggested in McCall and Irizarry (2011). If the batch structure is assumed to be known, batch effects can often be addressed by simple correction schemes such as median centering or by more advanced methods such as the empirical Bayes method *ComBat* (Johnson *et al.*, 2007). The latter technique is especially suitable for correcting data with small batch sizes and the incorporation of clinical covariates. When the batch vector is not known *Surrogate variable analysis* (SVA), a latent variable approach proposed by Leek and Storey (2007) provides an interesting alternative.

We extensively experimented with ComBat, SVA and variants of these to remove any of the possibly present batch effects in our data. In controlled experiments in which the batch vector was known, quantification of the batch effects before and after correction was based on the number of differentially expressed genes between batches, e.g. using *limma* (Smyth, 2005) or on Principle Variance Component Analysis (Scherer, 2009). Unfortunately, in most cases SVA, the more advanced method, failed to distinguish between relevant biological variation and unwanted technical variation, or the correction scheme itself clearly introduced unwanted variation. SVA works by decomposing a residual matrix via singular value decomposition or independent component analysis (Hyvärinen and Oja, 2000), to determine the latent variables, i.e. the *surrogate variables* (SVs), which explain systematic variation. The residual matrix, however, is obtained *after* the signal due to the primary variables of interest, i.e. known relevant biological variables, has been removed from the data. A critical step in SVA applications is, therefore, the correct specification of the model used to remove the signal of the primary variables. In case of a misspecified model, the residual matrix still contains relevant biological information, which in turn leads to true biological signal being misinterpreted as variation due to a confounder. This may be partly alleviated by performing a filter step on the SVs, i.e. by only selecting an SV if it correlates significantly with a known confounder, and does so more strongly than with any of the biological variables of interest (Teschendorff *et al.*, 2011). Although intriguing, it is hard not to think of Baron Münchhausen in this context, who allegedly pulled himself and his horse out of a swamp by his own hair. First, confounders are to a large degree unknown, which is precisely why we perform SVA. Second, one of the major goals of expression profiling is to identify *new* biological insights from the data, e.g. a refined set of breast cancer subtypes.

Therefore, it can be argued that the variables needed to perform the filter step are often not known at that stage. Given the difficulties in applying SVA to breast cancer data, we did not use these types of correction schemes and, when appropriate, relied on the simpler alternative of median centering.

Simple methods like median centering, however, are not always appropriate. A particular problem arises when batches are confounded with biologically relevant variables, e.g. a subtype vector. In the extreme case when all samples in a batch are of the same subtype, e.g. ER+, after batch correction by methods such as ComBat or median centering, approximately half of the samples in the batch will appear to have ER- characteristics. Especially processing time-based batches (Figure 6.1) are sensitive to this problem, since due to their often relatively small size the subtype distribution of their samples might be skewed towards one particular subtype and might not reflect the general population of breast cancers. One option is to omit the batch correction step entirely. Another approach would be to solve the problem during normalization, e.g. by using techniques such as frozen RMA (McCall *et al.*, 2010) (Chapter 3). Even though frozen RMA performed reasonably well in our experiments, in most cases it was clear that additional corrections were needed when combining data from distinct studies i.e. using a single reference distribution does often not fully remove between-study batch effects (data not shown).

6.2 Evaluation

In this dissertation breast cancer event prediction performance was mainly evaluated based on performance measures such as the balanced accuracy rate (BAR) and Area Under Receiver Operating Curve (AUC). These represent frequently applied performance measures in *machine learning*. Even though the former performance measures enjoy a wide-spread use, from a clinical perspective additional constraints on the evaluation process are desirable. For instance, regardless of the overall performance in terms of BAR or AUC, one could argue that a model is only clinically relevant if its sensitivity or negative predicted value is above a minimum threshold, say 0.9. Furthermore, our predictor evaluation was based on binary class labels. These represent a dichotomization of the survival data, with a cutoff at five years, i.e. a case belongs to the poor prognosis class if it had an event within five years and belongs to the good prognosis class if it did not have an event within five years, with a minimal follow-up of five years. It can be argued that these schemes do not fully exploit the available data. Indeed, patients with a follow-up shorter than five years and without an event, do not have a properly defined class label in our set-up and therefore had to be removed. Furthermore, it leaves us with

the unsettling notion that the data often contains various borderline cases, who had an event either just before or just after the cutpoint. The experiments offered in this dissertation, however, are all of sufficiently high sample size to avoid a large influence due to these issues. A powerful alternative is to not dichotomize the data and perform survival analysis. These approaches, e.g. Cox proportional hazards models (Cox, 1972), allow one to easily incorporate clinical information on grade, tumour size or age. Such covariates, however, are often only (publicly) available for a limited subset of the data. Therefore, such analyses were not attempted.

6.3 Refining the intrinsic subtypes

In Chapter 4 we utilized the intrinsic subtypes in order to improve event prediction. The main hope was that a stratification by subtype would lead to more homogeneous cohorts on which it would be easier to construct predictors. Even though improvements were observed in overall performance, they were only modest. In Chapter 3 we have seen that samples can be assigned to subtypes in a reproducible way, however, the predictor type and selected feature set may have still have a considerable influence on the subtype assigned to a sample. Therefore, the groups we have used may not have been as homogeneous as previously thought. Furthermore, of the intrinsic subtypes, only the luminal A subtype is clearly associated with a better outcome compared to the others. From a prediction perspective, a differentiation by this subtype will likely create the largest advantage in performance. It has indeed been reported that signatures such as MammaPrint a.k.a. the 70-gene signature by van't Veer *et al.* (2002) or Oncotype DX (Paik *et al.*, 2004), do precisely this, i.e. they separate luminal A samples from the other samples (Fan *et al.*, 2006; Prat *et al.*, 2011).

Over time various efforts have been made to refine the four intrinsic subtypes, i.e. luminal, basal, HER2+ and normal like, introduced by Perou *et al.* (2000). However, it is important to stress that even after a decade of genome-wide gene expression profiling, no single dataset has been put forward, for which a broad panel of key researchers in breast cancer, share consensus with respect to the subtypes of individual cases. Note that the St. Gallen consensus criteria (Chapter 3), only represent a set of definitions of the subtypes for which there is consensus. It does, however, not imply consensus on methodologies, measurement techniques or datasets which can be used to devise predictors that can identify the intrinsic subtypes. It is even questionable if such a consensus will ever be reached. For instance, it has been repeatedly observed that proliferation in luminal samples forms a continuum (Weigelt *et al.*, 2010b). Therefore,

any division into luminal A and B is somewhat arbitrary. Furthermore, the normal-like subtype is no longer considered to represent a true breast cancer subtype[1], and has been replaced by the *claudin-low subtype* (Prat *et al.*, 2010). The latter subtype, however, has not (yet) been adopted to the same extent as the other intrinsic subtypes. The relation between the HER2+ subtype and ER status also remains unclear. Some clearly see the HER2+ subtype as a subtype of the ER- branch (Goldhirsch *et al.*, 2011), while in SCMs, for instance, HER2+ can be both ER+ and ER- (Haibe-Kains *et al.*, 2012). In contrast, Guedj *et al.* (2011) suggest to replace the HER2+ subtype by a luminal C and *molecular apocrine subtype*. Furthermore, an alternative normal-like subtype is suggested. Of special interest are predictive markers for patients with triple-negative tumours. These tumours do not express ER, HER2 or PGR and are associated with very poor outcomes. Unfortunately, at present for these patients limited treatment options are available (Perou, 2011). Recently Lehmann *et al.* (2011) suggested that triple negative tumours may be further divided into as many as six distinct breast cancer subtypes. These observations suggest that the molecular subtypes of breast cancer are not sufficiently well understood and need further refinements, based on additional research and other measurement modalities.

6.4 Next-generation breast cancer compendia

Because of the costs associated with manufacturing and processing microarrays, initial microarray studies were often limited in the number of samples that could be targeted. In order to increase sample size, researchers were often forced to combine data from different platforms. This, however, also increased the level of unwanted technical variation. Furthermore, the first generation of breast cancer compendia mostly focussed on a single modality i.e. mRNA-based gene expression. Lower manufacturing and processing costs and the continuous development of alternative high-throughput measurement techniques have opened the door for the next generation of breast cancer compendia. The following subsections describe various aspects by which these compendia differ from the first generation of breast cancer compendia as studied in this dissertation.

6.4.1 Sample size

Over time array-based studies have become substantially larger. Single studies containing over a thousand samples are no longer uncommon. Furthermore,

[1] The normal-like subtype is still used, however, mostly for quality control purposes (Parker *et al.*, 2009).

in modern experiments better care is taken to reduce the level of unwanted technical variation by limiting the number of different platforms targeting the same modality. The level of heterogeneity can be further controlled by modern tissue extraction techniques such as laser capture dissection (Espina *et al.*, 2006) or single cell techniques (de Souza, 2011). The combination of lower costs and the ability to study cancer in large consortia, has made it feasible to study multiple tissue samples from a single patient on a much lager scale than previously possible. This offers multiple advantages. For instance, researchers can obtain a better perspective on the level of heterogeneity observed within a single tumour, by considering multiple tissue samples from the same tumour, taken at different spatial locations. Furthermore, it allows one to study tissues taken from multiple sites within the same patient, as well as to track the development of a disease over time by taking samples at different time points.

6.4.2 Alternative high-throughput technologies

Since the advent of the gene expression microarray in the mid-nineties, various other high-throughput measurement techniques have become available. Examples of these include *single nucleotide polymorphism arrays*, *array comparative genomic hybridization* (aCGH) and more recently, *next-generation sequencing* (NGS). Of these NGS methods arguably represent the largest breakthrough (Schuster, 2007). The sequencing counterpart of microarray based mRNA gene expression profiling is known as RNA-Seq (Mortazavi *et al.*, 2008). RNA-Seq has various advantages over microarray-based gene expression profiling. These include a higher resolution, virtually no background signal and overall better reproducibility ('t Hoen *et al.*, 2013). Furthermore, sequencing methods allow us to identify novel transcripts and new splice variants of known transcripts. However, like microarrays new high-throughput sequencing-based approaches are still susceptible to technical and biological biases and systematic errors that impact downstream analyses (Taub *et al.*, 2010; Risso *et al.*, 2011; Leek *et al.*, 2010; Zheng *et al.*, 2011).

6.4.3 Multimodal compendia

In this dissertation we mainly investigated the relationship between mRNA gene expression levels and breast cancer. However, the alternative measurement techniques described in the previous subsection allow one to study a variety of alternative modalities, e.g. copy number variations, single point mutations and DNA methylation. It may be advantageous to study multiple modalities at once for the same individual. Multimodal compendia offer complementary perspectives on the same biological phenomena and can therefore offer a more comprehensive understanding of the underlying biology than offered by any

single type of measurement. By now multimodal studies containing over a thousand samples are no longer uncommon and the first results of such studies have already been published. For instance, the Cancer Genome Atlas Network recently published a large multimodal study, the *Comprehensive molecular portraits of human breast tumours* (Cancer Genome Atlas Network, 2012). This study confirms the existence of the four main intrinsic subtypes, basal, HER2+, luminal A and luminal B. However, each of the four subtypes showed significant molecular heterogeneity. Even more heterogeneity was reported by the Molecular Taxonomy of Breast Cancer International Consortium (METABRIC) (Curtis *et al.*, 2012), based on copy number and gene expression data from nearly 2,000 patients. Based on a unsupervised analysis the authors suggest there may be at least 10 distinct breast cancer subtypes. Interestingly, the refinements include a very high-risk group in luminal breast cancers and a low-risk group in basal-like breast cancers. In the future multimodal views on breast cancer subtypes may lead to better definitions and characterizations of molecular breast cancer subtypes. In turn this may lead to improved prediction of survival and treatment response compared the performance achievable on microarray gene expression data alone. Evaluation protocols such as the one described in Chapter 4, are not specific to microarrays and can easily be applied to data from multimodal studies as well.

6.5 Microarray breast cancer profiling: success or failure?

At the time of their inception, microarrays were hailed as the new dawn in cancer biology and oncology practice (Colombo *et al.*, 2011). It is safe to say that after a decade of microarray breast cancer gene expression profiling, the intrinsic subtypes or first generation signatures such as the 70-gene profile by van't Veer *et al.* (2002), cannot be considered *mainstream* clinical techniques yet. Even though they are useful, they serve at most as complements to standard clinicopathological variables and not as replacements. In centralized assessments of clinicopathological variables with standardized methods, their added value has even been questioned (Cuzick *et al.*, 2011). Microarrays, however, have undeniably enhanced our understanding and acceptance of the inherent heterogeneity of breast cancer on a molecular level, even though this heterogeneity is currently only partly understood. It has become clear that most first-generation signatures merely represent substitutes for proliferation (Wirapati *et al.*, 2008). Furthermore, these markers have only limited prognostic value in ER- breast cancers. Second-generation signatures, i.e. subtype-specific signatures, such as the immune response signature by Teschendorff *et al.* (2007), or the GENIUS methodology by Haibe-Kains *et al.*

(2010), provide improvements of prognostic performance in ER- breast cancers. However, their performance is still not good enough for clinical purposes. One may therefore argue that microarrays have failed to live up to their promise. On the other hand, expectations may also simply have been too high. Regardless of the view one has on the contribution of microarrays, any measurement technique is bound to become obsolete given enough time. With the advent of next-generation sequencing, microarrays are unlikely to escape this fate. If anything, however, microarrays have helped to mature the field of bioinformatics and high-throughput gene expression profiling related techniques in general. Furthermore, they have laid a solid basis for the development of third-generation signatures, i.e. multimodal signatures. With all these new exciting measurement techniques and modalities, surely this time around cancer will be solved within a few years, or won't it?

SUMMARY

Microarrays offer biologists an exciting tool that allows the simultaneous assessment of gene expression levels for thousands of genes at once. At the time of their inception, microarrays were hailed as the new dawn in cancer biology and oncology practice with the hope that within a decade diseases like breast cancer would be solved. Various high-profile publications showed the immense potential of this technique in breast cancer event prediction and breast cancer subtyping. From these studies it became clear that breast cancer at the molecular level is not a single disease, but comprises a heterogeneous set of subtypes associated with clear differences in gene expression patterns and clinical outcomes. However, as microarrays became more popular, it became apparent that the accurate analysis and interpretation of microarray data provided a plethora of unique challenges. From a biological, as well as a technical perspective microarray data is complex, while the high feature-to-sample ratio associated with microarray studies rendered many classic statistical procedures useless. To make matters worse, various publications emerged that showed severe stability problems in the model fits of early pilot studies and showed that these studies were often overly optimistic. As a result the reliability of microarray based experiments in general was openly questioned. Given the multitude of different factors which may or may not influence results it is clear that a proper evaluation of microarray breast cancer profiling is both crucial and challenging.

This dissertation provides a number of carefully devised protocols, by which the influence of important sources of variation can be isolated, controlled and/or explicitly quantified, even in the absence of a gold standard. Instead of applying these protocols to data from small spike-in or dilution studies, they were applied to a large collection of real life breast cancer datasets of consider-

able size. Furthermore, we extensively studied breast cancer subtyping and the evaluation of subtype-specific predictors constructed on these, from both a practical and a theoretical perspective. This work shows that the evaluation of subtype-specific event prediction, based on divide and conquer schemes brings various new statistical challenges. For a variety of frequently encountered performance measures from machine learning several decompositions of the overall performance into subtype-specific performances are provided which show that the relation between subtype-specific and overall performance can be highly complex and counterintuitive. Furthermore, the experiments in this dissertation show that with modern processing techniques and a standardized approach it is possible to construct extremely stable subtyping schemes. However, the selected approach has a strong impact on the obtained results, suggesting that a stringent standardization of the methodologies used for subtyping is not sufficient for the consistent assignment of subtypes to individual patient samples. From these findings we conclude that the molecular subtypes of breast cancer are not sufficiently well understood and need further refinements.

SAMENVATTING

Microarrays bieden biologen de mogelijkheid om de expressieniveaus van duizenden genen tegelijkertijd te meten. Bij hun introductie werd de verwachting uitgesproken dat microarrays zouden leiden tot een beter begrip van de biologie/pathofysiologie van bijvoorbeeld kanker en revolutionaire toepassingen in de klinische praktijk, met de hoop dat binnen tien jaar ziektes zoals borstkanker opgelost zouden zijn. Verschillende vooraanstaande studies toonden al vroeg het immense potentieel van deze techniek aan bij het subtyperen van borstkanker en het voorspellen van metastasering. Uit deze studies bleek verder dat borstkanker op moleculair niveau niet één ziekte is, maar een verzameling heterogene subtypes geassocieerd met verschillende genexpressiepatronen en klinische uitkomsten. Echter, toen de populariteit van microarrays groeide werd duidelijk dat de nauwkeurige analyse en interpretatie van microarraydata vele unieke uitdagingen bevatte. Vanuit zowel biologisch als technisch oogpunt is microarraydata complex. De vele gemeten variabelen en het vaak beperkte aantal proefpersonen maakt directe toepassing van veel klassieke statistische methoden onmogelijk. Nog ernstiger is het feit dat de resultaten van diverse initiële studies zeer gevoelig bleken voor relatief kleine veranderingen in de data en dat hun conclusies vaak te optimistisch waren. Door deze bevindingen werd de betrouwbaarheid van microarray-experimenten en daaruit getrokken conclusies in het algemeen openlijk in twijfel getrokken. Uit het grote aantal factoren die mogelijkerwijs invloed uitoefenen op microarraydata blijkt duidelijk dat een juiste evaluatie van microarray-experimenten in de context van borstkankeronderzoek noodzakelijk, maar tegelijkertijd ook zeer uitdagend is.

Dit proefschrift beschrijft een aantal zorgvuldig ontworpen protocollen waarmee de invloed van belangrijke bronnen van variatie in microarray-experimenten

kan worden geïsoleerd, gecontroleerd en expliciet gekwantificeerd, zelfs bij het ontbreken van een gouden standaard. Hier worden deze protocollen toegepast op grote collecties microarraydata gemeten bij borstkankerpatiënten. Verder bestuderen we hoe borstkanker in verschillende subtypes kan worden opgedeeld op basis van genexpressiedata en onderzoeken we de relatie tussen subtypering en classificatie vanuit zowel praktisch als theoretisch perspectief. Dit werk toont aan dat de evaluatie van subtype-specifieke classificatiemodellen gebaseerd op een 'verdeel en heers' strategie diverse nieuwe statistische uitdagingen bevat. Voor een aantal in de patroonherkenning veelgebruikte uitkomstmaten biedt dit werk een exacte beschrijving van de totale nauwkeurigheid in termen van de subtype-specifieke nauwkeurigheden. Deze decomposities laten duidelijk zien dat de relatie tussen totale en subtype-specifieke nauwkeurigheid zeer complex en contra-intuïtief kan zijn. Verder laten de experimenten in dit proefschrift zien dat het met moderne pre-processing methodes en gestandaardiseerde benaderingen mogelijk is om subtypes zeer stabiel te detecteren. Echter, dit werk laat ook zien dat de gekozen aanpak de verkregen resultaten sterk beïnvloedt. Dit suggereert dat een strenge standaardisatie van de gebruikte methodes voor subtypering niet voldoende is voor een consistente toewijzing van subtypes aan individuele patiënten. Hieruit kan worden geconcludeerd dat de moleculaire subtypes van borstkanker nog niet voldoende goed begrepen zijn en verdere verfijningen behoeven.

ACKNOWLEDGMENTS

I would like to thank my promotor Marcel Reinders and copromotor Perry Moerland for their long term commitment and support. It was a very demanding, emotional experience, but ultimately a long set of welcome lessons in statistics, machine learning, computer science, biology and life in general. I am extremely thankful for all the time and extra hours that were spent by you on this research and for everything I was able to learn from you.

I would also like to thank Emile Aarts and Wim Verhaegh at Philips Research Laboratories for instigating this research and for their guidance during the initial stages of this work and thank Philips in general as the main sponsor of this work. In addition, I thank all my co-authors, especially René van den Ham, David Tax and Marc Uetz. I express my gratitude to my former colleagues at Philips; Paul van de Wiel, Bart Bakker, Jan Korst, Ellen Elfrink, Gaby Petznick, Dennis Merkle, Ruth de Boer, Nevenka Dimitrova, Ronald van Driel, Serge Vrijaldenhoven, Jurgen Rusch and all other former group members. In addition, I would like to thank Franklin Schuling and his wife for welcoming me to their home and making it possible to continue my PhD research in the United States of America in a time when all hope seemed lost. Furthermore, I thank Antoine van Kampen en Koos Zwinderman for their support and for providing a great working place at the Academic Medical Center in Amsterdam. In addition, I thank Benjamin Haibe-Kains for his work on the R genefu package, which was extensively used in many experiments described in this dissertation.

Most PhDs have their ups and downs and the same applies to this work. In those times family and friends are what is needed most. I want to thank my parents, Frits Sontrop and Maria Sontrop-Jongmans and my brother David

Sontrop for their emotional and financial support. I thank Gerry Satijn for proofreading this dissertation. Special thanks go to Annette Wijnhoven for her long term commitment, care and incredible patience. Without her help this dissertation would not have been possible. In addition, I thank all my family and friends, especially Pieter van der Horn, Xander van Scherpenzeel and Rimco van Rooij.

Finally, thanks to all my colleagues at Friss, where I found a new home. Special thanks go to Christian van Leeuwen, Roos Tji and Jeroen Morrenhof for allowing me to spend time on finishing this dissertation and for making me part of a great team.

Thank you all!

CURRICULUM VITAE

Herman M.J. Sontrop was born in Nijmegen on October 7, 1975. He spent most of his childhood in the Maastricht area. After completing his preliminary studies (VWO) at the Jeanne d'Arc college Maastricht, Herman obtained an MSc degree in Econometrics at the University of Maastricht. During his Masters he specialized in combinatorial optimization and in particular in solving complex vehicle routing problems by means of advanced meta-heuristic optimization schemes. His thesis was written at the Centre for Quantitative Methods (CQM) in Eindhoven. It was at CQM where he met prof. dr. ir. Emile Aarts, who asked him to join Philips. From 2005 till 2011 Herman worked for Philips Research Laboratories at the High Tech Campus in Eindhoven. At Philips Research Laboratories he worked in the area of bioinformatics with a specialization in gene-expression-based breast cancer subtyping and event prediction. In a later stage he joined forces with the Technical University of Delft and the Academic Medical Center (AMC) in Amsterdam where the majority of his PhD related work was written under supervision of prof. dr. ir. Marcel Reinders and dr. ir. Perry Moerland. In 2012 he started working for Friss, a fraud, risk and compliancy management firm based in Utrecht. Here he leads an analytic group which aims to improve risk assessment for insurers by using state-of-the-art machine learning techniques and implementations based on the R language for statistical computing. In addition, Herman highly enjoys working on advanced interactive data visualization techniques for the web.

PUBLICATIONS

Sontrop, H., Van Der Horn, P., Uetz, M. (2005). Fast ejection chain algorithms for vehicle routing with time windows (pp. 78-89). Springer Berlin Heidelberg.

Aarts, E., van der Horn, P., Korst, J., Michiels, W., **Sontrop, H.** (2006). Simulated Annealing. In Metaheuristic Procedures for Training Neutral Networks (pp. 37-52). Springer US.

Sontrop, H., van den Ham, R., Moerland, P., Reinders, M., Verhaegh, W. (2009, May). A sensitivity analysis of microarray feature selection and classification under measurement noise. In Genomic Signal Processing and Statistics, 2009. GENSIPS 2009. IEEE International Workshop on (pp. 1-4). IEEE.

Sontrop, H., Moerland, P., van den Ham, R., Reinders, M., Verhaegh, W. (2009). A comprehensive sensitivity analysis of microarray breast cancer classification under feature variability. BMC Bioinformatics, 10(1), 389.

Sontrop, H., Verhaegh, W., van den Ham, R., Reinders, M., Moerland, P. (2010, November). Subtype specific breast cancer event prediction. In Genomic Signal Processing and Statistics (GENSIPS), 2010 IEEE International Workshop on (pp. 1-4). IEEE.

Sontrop, H., Verhaegh, W., Reinders, M., Moerland, P. (2011). An evaluation protocol for subtype-specific breast cancer event prediction. PLoS ONE, 6(7), e21681.

Tax, D., **Sontrop, H.**, Reinders, M., Moerland, P. (2014). The Effect of Aggregating Subtype Performances Depends Strongly on the Performance Measure Used. In Pattern Recognition (ICPR), 2014 22nd International Conference on (pp. 3720-3725). IEEE.

Sontrop, H., Reinders, M., Moerland, P. (2014). Breast cancer subtype predictors revisited: from consensus to concordance? Submitted.

REFERENCES

Aarts, E. and Korst, J. (1990). *Simulated annealing and boltzmann machines. A stochastic approach to combinatorial optimization and neural computing*. New York, NY; John Wiley and Sons Inc.

Aarts, E. E. H. and Lenstra, J. K. (1997). *Local search in combinatorial optimization*. Princeton University Press.

Affymetrix (2002). *Statistical Algorithms Reference Guide*. Affymetrix.

Alexe, G., Dalgin, G. S., Ramaswamy, R., DeLisi, C., and Bhanot, G. (2006). Data perturbation independent diagnosis and validation of breast cancer subtypes using clustering and patterns. *Cancer Informatics*, **2**, 243–274.

Alon, U., Barkai, N., Notterman, D., Gish, K., Ybarra, S., Mack, D., and Levine, A. (1999). Broad patterns of gene expression revealed by clustering analysis of tumor and normal colon tissues probed by oligonucleotide arrays. *Proceedings of the National Academy of Sciences*, **96**(12), 6745.

Amaratunga, D. and Cabrera, J. (2004). *Exploration and analysis of DNA microarray and protein array data*. John Wiley Hoboken, NJ.

Ambroise, C. and McLachlan, G. J. (2002). Selection bias in gene extraction on the basis of microarray gene-expression data. *Proceedings of the National Academy of Sciences*, **99**(10), 6562–6566.

Baldi, P., Brunak, S., Chauvin, Y., Andersen, C., and Nielsen, H. (2000). Assessing the accuracy of prediction algorithms for classification: an overview. *Bioinformatics*, **16**(5), 412.

Barrett, T., Troup, D. B., Wilhite, S. E., Ledoux, P., Rudnev, D., Evangelista, C., Kim, I. F., Soboleva, A., Tomashevsky, M., Marshall, K. A., Phillippy, K. H., Sherman, P. M., Muertter, R. N., and Edgar, R. (2009). NCBI GEO: archive for high-throughput functional genomic data. *Nucleic Acids Research*, **37**(Database issue), D885–890.

Blagus, R. and Lusa, L. (2010). Class prediction for high-dimensional class-imbalanced data. *BMC Bioinformatics*, **11**(1), 523.

Bolstad, B. (2004). *Low-level analysis of high-density oligonucleotide array data: background, normalization and summarization*. Ph.D. thesis, University of California.

Bolstad, B., Collin, F., Brettschneider, J., Simpson, K., Cope, L., Irizarry, R., and Speed, T. (2005). Quality assessment of affymetrix genechip data. *Bioinformatics and computational biology solutions using R and bioconductor*, pages 33–47.

Bos, P., Xiang, H., Nadal, C., Shu, W., Gomis, R., Nguyen, D., Minn, A., van de Vijver, M., Gerald, W., Foekens, J., *et al.* (2009). Genes that mediate breast cancer metastasis to the brain. *Nature*, **459**(7249), 1005–1009.

Bradley, A. (1997). The use of the area under the ROC curve in the evaluation of machine learning algorithms. *Pattern Recognition*, **30**(7), 1145–1159.

Brazma, A., Parkinson, H., Sarkans, U., Shojatalab, M., Vilo, J., Abeygunawardena, N., Holloway, E., Kapushesky, M., Kemmeren, P., Lara, G., *et al.* (2003). ArrayExpress–a public repository for microarray gene expression data at the EBI. *Nucleic Acids Research*, **31**(1), 68–71.

Breiman, L. (2001). Random forests. *Machine learning*, **45**(1), 5–32.

Cancer Genome Atlas Network (2012). Comprehensive molecular portraits of human breast tumours. *Nature*, (490), 6170.

Causton, H., Quackenbush, J., and Brazma, A. (2009). *Microarray gene expression data analysis: a beginner's guide*. Wiley-Blackwell.

Chen, J., Hsueh, H., Delongchamp, R., Lin, C., and Tsai, C. (2007). Reproducibility of microarray data: a further analysis of microarray quality control (MAQC) data. *BMC Bioinformatics*, **8**(1), 412.

Chin, K., DeVries, S., Fridlyand, J., Spellman, P., Roydasgupta, R., Kuo, W., Lapuk, A., Neve, R., Qian, Z., Ryder, T., *et al.* (2006). Genomic and transcriptional aberrations linked to breast cancer pathophysiologies. *Cancer Cell*, **10**(6), 529–541.

Cohen, J. *et al.* (1960). A coefficient of agreement for nominal scales. *Educational and psychological measurement*, **20**(1), 37–46.

Collins, F. S., Morgan, M., and Patrinos, A. (2003). The human genome project: lessons from large-scale biology. *Science*, **300**(5617), 286–290.

Colombo, P., Milanezi, F., Weigelt, B., and Reis-Filho, J. (2011). Microarrays in the 2010s: the contribution of microarray-based gene expression profiling to breast cancer classification, prognostication and prediction. *Breast Cancer Research*, **13**(3), 212.

Cope, L., Irizarry, R., Jaffee, H., Wu, Z., and Speed, T. (2004). A benchmark for Affymetrix GeneChip expression measures. *Bioinformatics*, **20**(3), 323–331.

Copois, V., Bibeau, F., Bascoul-Mollevi, C., Salvetat, N., Chalbos, P., Bareil, C., Candeil, L., Fraslon, C., Conseiller, E., Granci, V., *et al.* (2007). Impact of rna degradation on gene expression profiles: assessment of different methods to reliably determine rna quality. *Journal of Biotechnology*, **127**(4), 549–559.

Cortes, C. and Mohri, M. (2004). AUC optimization vs. error rate minimization. In S. Thrun, L. Saul, and B. Schölkopf, editors, *Advances in Neural Information Processing Systems 16*. MIT Press, Cambridge, MA.

Cox, D. R. (1972). Regression models and life-tables. *Journal of the Royal Statistical Society. Series B (Methodological)*, pages 187–220.

Curtis, C., Shah, S. P., Chin, S.-F., Turashvili, G., Rueda, O. M., Dunning, M. J., Speed, D., Lynch, A. G., Samarajiwa, S., Yuan, Y., *et al.* (2012). The genomic and transcriptomic architecture of 2,000 breast tumours reveals novel subgroups. *Nature*, **486**(7403), 346–352.

Cuzick, J., Dowsett, M., Pineda, S., Wale, C., Salter, J., Quinn, E., Zabaglo, L., Mallon, E., Green, A. R., Ellis, I. O., *et al.* (2011). Prognostic value of a combined estrogen receptor, progesterone receptor, ki-67, and human epidermal growth factor receptor 2 immunohistochemical score and comparison with the genomic health recurrence score in early breast cancer. *Journal of Clinical Oncology*, **29**(32), 4273–4278.

de Souza, N. (2011). Single-cell methods. *Nature Methods*, **9**(1), 35–35.

Dedeurwaerder, S., Desmedt, C., Calonne, E., Singha, S., Haibe-Kains, B., Defrance, M., Michiels, S., Volkmar, M., Deplus, R., Luciani, J., *et al.* (2011). Dna methylation profiling reveals a predominant immune component in breast cancers. *EMBO Molecular Medicine*.

Desmedt, C., Piette, F., Loi, S., Wang, Y., Lallemand, F., Haibe-Kains, B., Viale, G., Delorenzi, M., Zhang, Y., d'Assignies, M., *et al.* (2007). Strong time dependence of the 76-gene prognostic signature for node-negative breast cancer patients in the transbig multicenter independent validation series. *Clinical cancer research*, **13**(11), 3207.

Desmedt, C., Haibe-Kains, B., Wirapati, P., Buyse, M., Larsimont, D., Bontempi, G., Delorenzi, M., Piccart, M., and Sotiriou, C. (2008). Biological processes associated with breast cancer clinical outcome depend on the molecular subtypes. *Clinical Cancer Research*, **14**(16), 5158.

Duda, R., Hart, P., and Stork, D. (2001). *Pattern classification*. Wiley New York.

Dudoit, S., Fridlyand, J., and Speed, T. (2002a). Comparison of discrimination methods for the classification of tumors using gene expression data. *Journal of the American Statistical Association*, **97**(457), 77–87.

Dudoit, S., Yang, Y. H., Callow, M. J., and Speed, T. P. (2002b). Statistical methods for identifying differentially expressed genes in replicated cdna microarray experiments. *Statistica sinica*, **12**(1), 111–140.

Edgar, R., Domrachev, M., and Lash, A. (2002). Gene Expression Omnibus: NCBI gene expression and hybridization array data repository. *Nucleic acids research*, **30**(1), 207.

Ein-Dor, L., Kela, I., Getz, G., Givol, D., and Domany, E. (2005). Outcome signature genes in breast cancer: is there a unique set? *Bioinformatics*, **21**(2), 171–178.

Espina, V., Wulfkuhle, J. D., Calvert, V. S., VanMeter, A., Zhou, W., Coukos, G., Geho, D. H., Petricoin, E. F., and Liotta, L. A. (2006). Laser-capture microdissection. *Nature Protocols*, **1**(2), 586–603.

Fan, C., Oh, D. S., Wessels, L., Weigelt, B., Nuyten, D. S., Nobel, A. B., van't Veer, L. J., and Perou, C. M. (2006). Concordance among gene-expression–based predictors for breast cancer. *New England Journal of Medicine*, **355**(6), 560–569.

Fare, T., Coffey, E., Dai, H., He, Y., Kessler, D., Kilian, K., Koch, J., LeProust, E., Marton, M., Meyer, M., *et al.* (2003). Effects of atmospheric ozone on microarray data quality. *Analytical chemistry Washington DC*, **75**(17), 4672–4675.

Farmer, P., Bonnefoi, H., Becette, V., Tubiana-Hulin, M., Fumoleau, P., Larsimont, D., MacGrogan, G., Bergh, J., Cameron, D., Goldstein, D., *et al.* (2005). Identification of molecular apocrine breast tumours by microarray analysis. *Breast Cancer Research*, **7**, 1–1.

Foekens, J., Atkins, D., Zhang, Y., Sweep, F., Harbeck, N., Paradiso, A., Cufer, T., Sieuwerts, A., Talantov, D., Span, P., *et al.* (2006). Multicenter validation of a gene expression-based prognostic signature in lymph node-negative primary breast cancer. *Journal of Clinical Oncology*, **24**(11), 1665.

Friedman, J., Hastie, T., and Tibshirani, R. (2001). *The elements of statistical learning*, volume 1. Springer Series in Statistics.

Gautier, L., Cope, L., Bolstad, B., and Irizarry, R. (2004). affy - analysis of Affymetrix GeneChip data at the probe level. *Bioinformatics*, **20**(3), 307–315.

Geman, D. *et al.* (2004). Classifying gene expression profiles from pairwise mRNA comparisons. *Statistical applications in genetics and molecular biology*, **3**, 19.

Gentleman, R., Carey, V., Bates, D., Bolstad, B., Dettling, M., Dudoit, S., Ellis, B., Gautier, L., Ge, Y., Gentry, J., *et al.* (2004). Bioconductor: open software development for computational biology and bioinformatics. *Genome Biology*, **5**(10), R80.

Glover, F., Laguna, M., *et al.* (1997). *Tabu search*, volume 22. Springer.

Gohlmann, H. and Talloen, W. (2010). *Gene expression studies using Affymetrix microarrays*. Chapman and Hall/CRC.

Goldberg, D. E. and Holland, J. H. (1988). Genetic algorithms and machine learning. *Machine learning*, **3**(2), 95–99.

Goldhirsch, A., Wood, W., Coates, A., Gelber, R., Thürlimann, B., Senn, H., *et al.* (2011). Strategies for subtypes - dealing with the diversity of breast cancer: highlights of the St Gallen International Expert Consonsensus on the Primary Therapy of Early Breast Cancer 2011. *Annals of Oncology*, **22**(8), 1736–1747.

Golub, T., Slonim, D., Tamayo, P., Huard, C., Gaasenbeek, M., Mesirov, J., Coller, H., Loh, M., Downing, J., Caligiuri, M., *et al.* (1999). Molecular classification of cancer: class discovery and class prediction by gene expression monitoring. *Science*, **286**(5439), 531.

Guedj, M., Marisa, L., De Reynies, A., Orsetti, B., Schiappa, R., Bibeau, F., MacGrogan, G., Lerebours, F., Finetti, P., Longy, M., Bertheau, P., *et al.* (2011). Gene expression profiling for molecular characterization of inflammatory breast cancer and prediction of response to chemotherapy. *Oncogene*.

Guiu, S., Michiels, S., Andre, F., Cortes, J., Denkert, C., Di Leo, A., Hennessy, B., Sorlie, T., Sotiriou, C., Turner, N., *et al.* (2012). Molecular subclasses of breast cancer: how do we define them? the IMPAKT 2012 Working Group Statement. *Annals of Oncology*, **23**(12), 2997–3006.

Guo, Y., Hastie, T., and Tibshirani, R. (2007). Regularized linear discriminant analysis and its application in microarrays. *Biostatistics*, **8**(1), 86–100.

Haibe-Kains, B. (2009). *Identification and assessment of gene signatures in human breast cancer*. Ph.D. thesis, University Libre de Bruxelles, Bioinformatics Department.

Haibe-Kains, B., Desmedt, C., Rothé, F., Piccart, M., Sotiriou, C., and Bontempi, G. (2010). A fuzzy gene expression-based computational approach improves breast cancer prognostication. *Genome Biology*, **11**(2), R18.

Haibe-Kains, B., Desmedt, C., Loi, S., Culhane, A., Bontempi, G., Quackenbush, J., and Sotiriou, C. (2012). A three-gene model to robustly identify breast cancer molecular subtypes. *Journal of the National Cancer Institute*.

Hastie, T. and Tibshirani, R. (2004). Efficient quadratic regularization for expression arrays. *Biostatistics*, **5**(3), 329–340.

Hatzis, C., Sun, H., Yao, H., Hubbard, R. E., Meric-Bernstam, F., Babiera, G. V., Wu, Y., Pusztai, L., and Symmans, W. F. (2011). Effects of tissue handling on rna integrity and microarray measurements from resected breast cancers. *Journal of the National Cancer Institute*, **103**(24), 1871–1883.

He, H. and Garcia, E. (2009). Learning from imbalanced data. *IEEE Transactions on Knowledge and Data Engineering*, **21**(9), 1263–1284.

Hoffmann, R., Seidl, T., and Dugas, M. (2002). Profound effect of normalization on detection of differentially expressed genes in oligonucleotide microarray data analysis. *Genome Biol*, **3**(7), 0033–1.

Hu, Z., Fan, C., Oh, D., Marron, J., He, X., Qaqish, B., Livasy, C., Carey, L., Reynolds, E., Dressler, L., *et al.* (2006). The molecular portraits of breast tumors are conserved across microarray platforms. *BMC genomics*, **7**(1), 96.

Hyvärinen, A. and Oja, E. (2000). Independent component analysis: algorithms and applications. *Neural networks*, **13**(4), 411–430.

Ihaka, R. and Gentleman, R. (1996). R: a language for data analysis and graphics. *Journal of Computational and Graphical Statistics*, **5**(3), 299–314.

Irizarry, R., Hobbs, B., Collin, F., Beazer-Barclay, Y., Antonellis, K., Scherf, U., and Speed, T. (2003). Exploration, normalization, and summaries of high density oligonucleotide array probe level data. *Biostatistics*, **4**(2), 249–264.

Irizarry, R., Wu, Z., and Jaffee, H. (2006). Comparison of Affymetrix GeneChip expression measures. *Bioinformatics*, **22**(7), 789–794.

Japkowicz, N. and Stephen, S. (2002). The class imbalance problem: A systematic study. *Intelligent Data Analysis*, **6**, 429–449.

Johnson, W., Li, C., and Rabinovic, A. (2007). Adjusting batch effects in microarray expression data using empirical Bayes methods. *Biostatistics*, **8**(1), 118.

Kao, K., Chang, K., Hsu, H., and Huang, A. (2011). Correlation of microarray-based breast cancer molecular subtypes and clinical outcomes: implications for treatment optimization. *BMC cancer*, **11**(1), 143.

Kapp, A., Jeffrey, S., Langerod, A., Borresen-Dale, A., Han, W., Noh, D., Bukholm, I., Nicolau, M., Brown, P., and Tibshirani, R. (2006). Discovery and validation of breast cancer subtypes. *BMC genomics*, **7**(231), 1471–2164.

Kauffmann, A. and Huber, W. (2010). Microarray data quality control improves the detection of differentially expressed genes. *Genomics*, **95**(3), 138–142.

Kauffmann, A., Gentleman, R., and Huber, W. (2009). arrayQualityMetrics–a bioconductor package for quality assessment of microarray data. *Bioinformatics*, **25**(3), 415.

Kim, S. (2009). Effects of sample size on robustness and prediction accuracy of a prognostic gene signature. *BMC Bioinformatics*, **10**(1), 147.

Klebanov, L. and Yakovlev, A. (2007). How high is the level of technical noise in microarray data. *Biology Direct*, **2**(1), 9.

Kohavi, R. (1995). A study of cross-validation and bootstrap for accuracy estimation and model selection. In *Proceedings of the 15th International Joint Conference on Artificial Intelligence*, pages 1137–1143.

Kreil, D. and Russell, R. (2005). There is no silver bullet - a guide to low-level data transforms and normalisation methods for microarray data. *Briefings in Bioinformatics*, **6**(1), 86–97.

Lai, C., Reinders, M., van't Veer, L., Wessels, L., *et al.* (2006). A comparison of univariate and multivariate gene selection techniques for classification of cancer datasets. *BMC Bioinformatics*, **7**(1), 235.

Leek, J. and Storey, J. (2007). Capturing heterogeneity in gene expression studies by surrogate variable analysis. *PLoS Genet*, **3**(9), e161.

Leek, J., Scharpf, R., Bravo, H., Simcha, D., Langmead, B., Johnson, W., Geman, D., Baggerly, K., and Irizarry, R. (2010). Tackling the widespread and critical impact of batch effects in high-throughput data. *Nature Reviews Genetics*, **11**(10), 733–739.

Lehmann, B. D., Bauer, J. A., Chen, X., Sanders, M. E., Chakravarthy, A. B., Shyr, Y., and Pietenpol, J. A. (2011). Identification of human triple-negative breast cancer subtypes and preclinical models for selection of targeted therapies. *The Journal of clinical investigation*, **121**(7), 2750.

Li, C. and Wong, W. (2001). Model-based analysis of oligonucleotide arrays: expression index computation and outlier detection. *Proceedings of the National Academy of Sciences*, **98**(1), 31–36.

Li, Y., Zou, L., Li, Q., Haibe-Kains, B., Tian, R., Li, Y., Desmedt, C., Sotiriou, C., Szallasi, Z., Iglehart, J., *et al.* (2010). Amplification of laptm4b and ywhaz contributes to chemotherapy resistance and recurrence of breast cancer. *Nature medicine*, **16**(2), 214–218.

Liang, P. (2007). MAQC papers over the cracks. *Nature Biotechnology*, **25**(1), 27–8.

Liu, X., Milo, M., Lawrence, N., and Rattray, M. (2005). A tractable probabilistic model for Affymetrix probe-level analysis across multiple chips. *Bioinformatics*, **21**(18), 3637–3644.

Liu, X., Milo, M., Lawrence, N., and Rattray, M. (2006). Probe-level measurement error improves accuracy in detecting differential gene expression. *Bioinformatics*, **22**(17), 2107–2113.

Lockhart, D. J., Dong, H., Byrne, M. C., Follettie, M. T., Gallo, M. V., Chee, M. S., Mittmann, M., Wang, C., Kobayashi, M., Norton, H., *et al.* (1996). Expression monitoring by hybridization to high-density oligonucleotide arrays. *Nature biotechnology*, **14**(13), 1675–1680.

Loi, S., Haibe-Kains, B., Desmedt, C., Lallemand, F., Tutt, A., Gillet, C., Ellis, P., Harris, A., Bergh, J., Foekens, J., *et al.* (2007). Definition of clinically distinct molecular subtypes in estrogen receptor–positive breast carcinomas through genomic grade. *Journal of clinical oncology*, **25**(10), 1239.

Lu, X., Lu, X., Wang, Z., Iglehart, J., Zhang, X., and Richardson, A. (2008). Predicting features of breast cancer with gene expression patterns. *Breast cancer research and treatment*, **108**(2), 191–201.

Lusa, L., McShane, L., Reid, J., De Cecco, L., Ambrogi, F., Biganzoli, E., Gariboldi, M., and Pierotti, M. (2007). Challenges in Projecting Clustering Results Across Gene Expression Profiling Datasets. *JNCI Journal of the National Cancer Institute*.

Mackay, A., Weigelt, B., Grigoriadis, A., Kreike, B., Tan, D., Dowsett, M., Ashworth, A., and Reis-Filho, J. (2011). Microarray-based class discovery for molecular classification of breast cancer: analysis of interobserver agreement. *Journal of the National Cancer Institute*, **103**(8), 662.

McCall, M. and Irizarry, R. (2008). Consolidated strategy for the analysis of microarray spike-in data. *Nucleic Acids Research*, **36**(17), e108.

McCall, M. and Irizarry, R. (2011). Thawing frozen robust multi-array analysis (frma). *BMC bioinformatics*, **12**(1), 369.

McCall, M., Bolstad, B., and Irizarry, R. (2010). Frozen robust multiarray analysis (frma). *Biostatistics*, **11**(2), 242.

McCall, M., Murakami, P., Lukk, M., Huber, W., and Irizarry, R. (2011). Assessments of affymetrix genechip microarray quality for laboratories and single samples. *BMC bioinformatics*, **12**(1), 137.

McClintick, J. and Edenberg, H. (2006). Effects of filtering by Present call on analysis of microarray experiments. *BMC Bioinformatics*, **7**(1), 49.

Medeiros, F., Rigl, C. T., Anderson, G. G., Becker, S. H., and Halling, K. C. (2007). Tissue handling for genome-wide expression analysis: a review of the issues, evidence, and opportunities. *Archives of pathology & laboratory medicine*, **131**(12), 1805–1816.

Michiels, S., Koscielny, S., and Hill, C. (2005). Prediction of cancer outcome with microarrays: a multiple random validation strategy. *The Lancet*, **365**(9458), 488–492.

Miller, L., Smeds, J., George, J., Vega, V., Vergara, L., Ploner, A., Pawitan, Y., Hall, P., Klaar, S., Liu, E., *et al.* (2005). An expression signature for p53 status in human breast cancer predicts mutation status, transcriptional effects, and patient survival. *Proceedings of the National Academy of Sciences*, **102**(38), 13550–5.

Milo, M., Fazeli, A., Niranjan, M., and Lawrence, N. (2003). A probabilistic model for the extraction of expression levels from oligonucleotide arrays. *Biochemical Society Transactions*, **31**, 1510–1512.

Minn, A., Gupta, G., Siegel, P., Bos, P., Shu, W., Giri, D., Viale, A., Olshen, A., Gerald, W., and Massagué, J. (2005). Genes that mediate breast cancer metastasis to lung. *Nature*, **436**, 518–524.

Mitchell, T. M. (1997). Machine learning. wcb.

Moore, D. (1996). *Statistics, concepts and controversies*. W.H. Freeman and Company, New York, 4th edition.

Mortazavi, A., Williams, B. A., McCue, K., Schaeffer, L., and Wold, B. (2008). Mapping and quantifying mammalian transcriptomes by rna-seq. *Nature methods*, **5**(7), 621–628.

Natrajan, R., Weigelt, B., Mackay, A., Geyer, F., Grigoriadis, A., Tan, D., Jones, C., Lord, C., Vatcheva, R., Rodriguez-Pinilla, S., *et al.* (2010). An integrative genomic and transcriptomic analysis reveals molecular pathways and networks regulated by copy number aberrations in basal-like, her2 and luminal cancers. *Breast cancer research and treatment*, **121**(3), 575–589.

Nielsen, T. O., Parker, J. S., Leung, S., Voduc, D., Ebbert, M., Vickery, T., Davies, S. R., Snider, J., Stijleman, I. J., Reed, J., *et al.* (2010). A comparison of PAM50 intrinsic subtyping with immunohistochemistry and clinical prognostic factors in tamoxifen-treated estrogen receptor–positive breast cancer. *Clinical Cancer Research*, **16**(21), 5222–5232.

Paik, S., Shak, S., Tang, G., Kim, C., Baker, J., Cronin, M., Baehner, F. L., Walker, M. G., Watson, D., Park, T., *et al.* (2004). A multigene assay to predict recurrence of tamoxifen-treated, node-negative breast cancer. *New England Journal of Medicine*, **351**(27), 2817–2826.

Parker, B., Günter, S., and Bedo, J. (2007). Stratification bias in low signal microarray studies. *BMC Bioinformatics*, **8**(1), 326.

Parker, J., Mullins, M., Cheang, M., Leung, S., Voduc, D., Vickery, T., Davies, S., Fauron, C., He, X., Hu, Z., *et al.* (2009). Supervised risk predictor of

breast cancer based on intrinsic subtypes. *Journal of Clinical Oncology*, **27**(8), 1160.

Parkinson, H., Sarkans, U., Shojatalab, M., Abeygunawardena, N., Contrino, S., Coulson, R., Farne, A., Garcia Lara, G., Holloway, E., Kapushesky, M., *et al.* (2005). ArrayExpressa public repository for microarray gene expression data at the EBI. *Nucleic acids research*, **33**(suppl 1), D553.

Pawitan, Y., Bjöhle, J., Amler, L., Borg, A., Egyhazi, S., Hall, P., Han, X., Holmberg, L., Huang, F., Klaar, S., *et al.* (2005). Gene expression profiling spares early breast cancer patients from adjuvant therapy: derived and validated in two population-based cohorts. *Breast Cancer Research*, **7**(6), R953–R964.

Pearson, R., Liu, X., Sanguinetti, G., Milo, M., Lawrence, N., and Rattray, M. (2009). puma: a Bioconductor package for Propagating Uncertainty in Microarray Analysis. *BMC Bioinformatics*, **10**(1), 211.

Perou, C., Sørlie, T., Eisen, M., van de Rijn, M., Jeffrey, S., Rees, C., Pollack, J., Ross, D., Johnsen, H., Akslen, L., *et al.* (2000). Molecular portraits of human breast tumours. *Nature*, **406**(6797), 747–752.

Perou, C., Parker, J., Prat, A., Ellis, M., and Bernard, P. (2010). Clinical implementation of the intrinsic subtypes of breast cancer. *Lancet Oncology*, **11**(8), 718–719.

Perou, C. M. (2011). Molecular stratification of triple-negative breast cancers. *The Oncologist*, **16**(Supplement 1), 61–70.

Popovici, V., Chen, W., Gallas, B., Hatzis, C., Shi, W., Samuelson, F., Nikolsky, Y., Tsyganova, M., Ishkin, A., Nikolskaya, T., *et al.* (2010). Effect of training-sample size and classification difficulty on the accuracy of genomic predictors. *Breast Cancer Res*, **12**(1), R5.

Prat, A., Parker, J., Karginova, O., Fan, C., Livasy, C., Herschkowitz, J., He, X., Perou, C., *et al.* (2010). Phenotypic and molecular characterization of the claudin-low intrinsic subtype of breast cancer. *Breast Cancer Res*, **12**(5), R68.

Prat, A., Ellis, M., and Perou, C. (2011). Practical implications of gene-expression-based assays for breast oncologists. *Nature Reviews Clinical Oncology*.

Pusztai, L., Mazouni, C., Anderson, K., Wu, Y., and Symmans, W. (2006). Molecular classification of breast cancer: limitations and potential. *The Oncologist*, **11**(8), 868–877.

Rattray, M., Liu, X., Sanguinetti, G., Milo, M., and Lawrence, N. (2006). Propagating uncertainty in microarray data analysis. *Briefings in Bioinformatics*, **7**(1), 37–47.

Reyal, F., van Vliet, M., Armstrong, N., Horlings, H., de Visser, K., Kok, M., Teschendorff, A., Mook, S., Caldas, C., Salmon, R., *et al.* (2008). A comprehensive analysis of prognostic signatures reveals the high predictive capacity of the proliferation, immune response and RNA splicing modules in breast cancer. *Breast Cancer Research*, **10**(6), R93.

Richardson, A., Wang, Z., De Nicolo, A., Lu, X., Brown, M., Miron, A., Liao, X., Iglehart, J., Livingston, D., and Ganesan, S. (2006). X chromosomal abnormalities in basal-like human breast cancer. *Cancer cell*, **9**(2), 121–132.

Risso, D., Schwartz, K., Sherlock, G., and Dudoit, S. (2011). Gc-content normalization for rna-seq data. *BMC bioinformatics*, **12**(1), 480.

Ritchie, M., Silver, J., Oshlack, A., Holmes, M., Diyagama, D., Holloway, A., and Smyth, G. (2007). A comparison of background correction methods for two-colour microarrays. *Bioinformatics*, **23**(20), 2700–2707.

Roepman, P., Wessels, L., Kettelarij, N., Kemmeren, P., Miles, A., Lijnzaad, P., Tilanus, M., Koole, R., Hordijk, G., van der Vliet, P., *et al.* (2005). An expression profile for diagnosis of lymph node metastases from primary head and neck squamous cell carcinomas. *Nature Genetics*, **37**(2), 182–186.

Roepman, P., Kemmeren, P., Wessels, L., Slootweg, P., and Holstege, F. (2006). Multiple robust signatures for detecting lymph node metastasis in head and neck cancer. *Cancer Research*, **66**(4), 2361–2366.

Rota, G. (1964). The number of partitions of a set. *American Mathematical Monthly*, **71**(5), 498–504.

Sabatier, R., Finetti, P., Cervera, N., Lambaudie, E., Esterni, B., Mamessier, E., Tallet, A., Chabannon, C., Extra, J., Jacquemier, J., *et al.* (2011). A gene expression signature identifies two prognostic subgroups of basal breast cancer. *Breast cancer research and treatment*, **126**(2), 407–420.

Sanguinetti, G., Milo, M., Rattray, M., and Lawrence, N. (2005). Accounting for probe-level noise in principal component analysis of microarray data. *Bioinformatics*, **21**(19), 3748–3754.

Schaffer, J. D., Janevski, A., and Simpson, M. R. (2005). A genetic algorithm approach for discovering diagnostic patterns in molecular measurement data. In *Computational Intelligence in Bioinformatics and Computational Biology,*

2005. CIBCB'05. Proceedings of the 2005 IEEE Symposium on, pages 1–8. IEEE.

Schena, M., Shalon, D., Davis, R. W., and Brown, P. O. (1995). Quantitative monitoring of gene expression patterns with a complementary dna microarray. *Science*, **270**(5235), 467–470.

Scherer, A. (2009). *Batch Effects and Noise in Microarray Experiments: Sources and Solutions*. Wiley.

Schmidt, M., Böhm, D., von Törne, C., Steiner, E., Puhl, A., Pilch, H., Lehr, H., Hengstler, J., Kölbl, H., and Gehrmann, M. (2008). The humoral immune system has a key prognostic impact in node-negative breast cancer. *Cancer research*, **68**(13), 5405.

Schuster, S. C. (2007). Next-generation sequencing transforms todays biology. *Nature*, **200**(8).

Shi, L., Reid, L., Jones, W., Shippy, R., Warrington, J., Baker, S., Collins, P., de Longueville, F., Kawasaki, E., and et al., L. K. (2006). The microarray quality control (maqc) project shows inter-and intraplatform reproducibility of gene expression measurements. *Nature Biotechnology*, **24**, 1151–1161.

Shi, L., Campbell, G., Jones, W., Campagne, F., Wen, Z., Walker, S., Su, Z., Chu, T., Goodsaid, F., Pusztai, L., *et al.* (2010). The microarray quality control (maqc)-ii study of common practices for the development and validation of microarray-based predictive models. *Nature biotechnology*, **28**(8), 827.

Shipp, M., Ross, K., Tamayo, P., Weng, A., Kutok, J., Aguiar, R., Gaasenbeek, M., Angelo, M., Reich, M., Pinkus, G., *et al.* (2002). Diffuse large B-cell lymphoma outcome prediction by gene-expression profiling and supervised machine learning. *Nature medicine*, **8**(1), 68–74.

Siegel, R., Ward, E., Brawley, O., and Jemal, A. (2011). Cancer statistics, 2011. *CA: a cancer journal for clinicians*, **61**(4), 212–236.

Simon, R., Radmacher, M. D., Dobbin, K., and McShane, L. M. (2003). Pitfalls in the use of dna microarray data for diagnostic and prognostic classification. *Journal of the National Cancer Institute*, **95**(1), 14–18.

Singh, D., Febbo, P., Ross, K., Jackson, D., Manola, J., Ladd, C., Tamayo, P., Renshaw, A., D'Amico, A., Richie, J., *et al.* (2002). Gene expression correlates of clinical prostate cancer behavior. *Cancer cell*, **1**(2), 203–209.

Smyth, G. (2005). Limma: linear models for microarray data. In R. Gentleman, V. Carey, S. Dudoit, R. Irizarry, W. Huber, and S. Dudoit, editors, *Bioinformatics and Computational Biology Solutions Using R and Bioconductor*, pages 397–420. Springer, New York.

Sontrop, H., Moerland, P., Van Den Ham, R., Reinders, M., and Verhaegh, W. (2009). A comprehensive sensitivity analysis of microarray breast cancer classification under feature variability. *BMC Bioinformatics*, **10**(1), 389.

Sontrop, H., Verhaegh, W., Reinders, M., and Moerland, P. (2011). An evaluation protocol for subtype-specific breast cancer event prediction. *PLoS ONE*, **6**(7), e21681.

Sørlie, T., Perou, C., Tibshirani, R., Aas, T., Geisler, S., Johnsen, H., Hastie, T., Eisen, M., Van De Rijn, M., Jeffrey, S., *et al.* (2001). Gene expression patterns of breast carcinomas distinguish tumor subclasses with clinical implications. *Proceedings of the National Academy of Sciences*, **98**(19), 10869.

Sørlie, T., Tibshirani, R., Parker, J., Hastie, T., Marron, J., Nobel, A., Deng, S., Johnsen, H., Pesich, R., Geisler, S., *et al.* (2003). Repeated observation of breast tumor subtypes in independent gene expression data sets. *Proceedings of the National Academy of Sciences*, **100**(14), 8418.

Sørlie, T., Borgan, E., Myhre, S., Vollan, H., Russnes, H., Zhao, X., Nilsen, G., Lingjærde, O., Børresen-Dale, A., and Rødland, E. (2010). The importance of gene-centring microarray data. *The Lancet Oncology*, **11**(8), 719–720.

Sotiriou, C. and Piccart, M. (2007). Taking gene-expression profiling to the clinic: when will molecular signatures become relevant to patient care? *Nature Reviews Cancer*, **7**(7), 545–553.

Sotiriou, C., Wirapati, P., Loi, S., Harris, A., Fox, S., Smeds, J., Nordgren, H., Farmer, P., Praz, V., Haibe-Kains, B., *et al.* (2006). Gene expression profiling in breast cancer: understanding the molecular basis of histologic grade to improve prognosis. *Journal of the National Cancer Institute*, **98**(4), 262.

Speed, T. (2003). *Statistical analysis of gene expression microarray data*, volume 11. Chapman and Hall/CRC.

Stafford, P. and Brun, M. (2007). Three methods for optimization of cross-laboratory and cross-platform microarray expression data. *Nucleic Acids Research*, **35**(10), e72.

Statnikov, A., Wang, L., and Aliferis, C. (2008). A comprehensive comparison of random forests and support vector machines for microarray-based cancer classification. *BMC Bioinformatics*, **9**(1), 319.

Symmans, W., Hatzis, C., Sotiriou, C., Andre, F., Peintinger, F., Regitnig, P., Daxenbichler, G., Desmedt, C., Domont, J., Marth, C., *et al.* (2010). Genomic index of sensitivity to endocrine therapy for breast cancer. *Journal of Clinical Oncology*, **28**(27), 4111.

't Hoen, Friedländer, M. R., Almlöf, J., Sammeth, M., Pulyakhina, I., Anvar, S. Y., Laros, J. F., Buermans, H. P., Karlberg, O., Brännvall, M., *et al.* (2013). Reproducibility of high-throughput mrna and small rna sequencing across laboratories. *Nature biotechnology*.

Tan, A., Naiman, D., Xu, L., Winslow, R., and Geman, D. (2005). Simple decision rules for classifying human cancers from gene expression profiles. *Bioinformatics*, **21**(20), 3896.

Taub, M. A., Corrada Bravo, H., and Irizarry, R. A. (2010). Overcoming bias and systematic errors in next generation sequencing data. *Genome Med*, **2**(12), 87.

Teschendorff, A., Miremadi, A., Pinder, S., Ellis, I., Caldas, C., *et al.* (2007). An immune response gene expression module identifies a good prognosis subtype in estrogen receptor negative breast cancer. *Genome Biol*, **8**(8), R157.

Teschendorff, A. E., Zhuang, J., and Widschwendter, M. (2011). Independent surrogate variable analysis to deconvolve confounding factors in large-scale microarray profiling studies. *Bioinformatics*, **27**(11), 1496–1505.

Tibshirani, R. and Walther, G. (2005). Cluster validation by prediction strength. *Journal of Computational and Graphical Statistics*, **14**(3), 511–528.

Tu, Y., Stolovitzky, G., and Klein, U. (2002). Quantitative noise analysis for gene expression microarray experiments. *Proceedings of the National Academy of Sciences*, **99**(22), 14031–14036.

van de Vijver, M., He, Y., van't Veer, L., Dai, H., Hart, A., Voskuil, D., Schreiber, G., Peterse, J., Roberts, C., Marton, M., *et al.* (2002). A Gene-Expression Signature as a Predictor of Survival in Breast Cancer. *New England Journal of Medicine*, **347**(25), 1999–2009.

van Rijsbergen, C. J. (1979). *Information Retrieval*. Butterwort, London, 2nd edition.

van Vliet, M., Reyal, F., Horlings, H., van de Vijver, M., Reinders, M., and Wessels, L. (2008). Pooling breast cancer datasets has a synergetic effect on classification performance and improves signature stability. *BMC Genomics*, **9**(1), 375.

van't Veer, L. and Bernards, R. (2008). Enabling personalized cancer medicine through analysis of gene-expression patterns. *Nature*, **452**(7187), 564.

van't Veer, L., Dai, H., van de Vijver, M., He, Y., Hart, A., Mao, M., Peterse, H., van der Kooy, K., Marton, M., Witteveen, A., *et al.* (2002). Gene expression profiling predicts clinical outcome of breast cancer. *Nature*, **415**(6871), 530–536.

Verhaak, R., Staal, F., Valk, P., Lowenberg, B., Reinders, M., and de Ridder, D. (2006). The effect of oligonucleotide microarray data pre-processing on the analysis of patient-cohort studies. *BMC Bioinformatics*, **7**(1), 105.

Wang, J., Wen, S., Symmans, W., Pusztai, L., and Coombes, K. (2009). The bimodality index: a criterion for discovering and ranking bimodal signatures from cancer gene expression profiling data. *Cancer informatics*, **7**, 199.

Wang, X., Markowetz, F., De Sousa E Melo, F., Medema, J. P., and Vermeulen, L. (2013). Dissecting cancer heterogeneity–an unsupervised classification approach. *The International Journal of Biochemistry & Cell Biology*, **45**(11), 2574–2579.

Wang, Y., Klijn, J., Zhang, Y., Sieuwerts, A., Look, M., Yang, F., Talantov, D., Timmermans, M., Meijer-van Gelder, M., Yu, J., *et al.* (2005). Gene-expression profiles to predict distant metastasis of lymph-node-negative primary breast cancer. *The Lancet*, **365**(9460), 671–679.

Watson, J. D. (1990). The human genome project: past, present, and future. *Science*, **248**(4951), 44–49.

Weigelt, B., Mackay, A., A'hern, R., Natrajan, R., Tan, D., Dowsett, M., Ashworth, A., and Reis-Filho, J. (2010a). Breast cancer molecular profiling with single sample predictors: a retrospective analysis. *The Lancet Oncology*, **11**(4), 339–349.

Weigelt, B., Baehner, F. L., and Reis-Filho, J. S. (2010b). The contribution of gene expression profiling to breast cancer classification, prognostication and prediction: a retrospective of the last decade. *The Journal of pathology*, **220**(2), 263–280.

Weigelt, B., Mackay, A., A'hern, R., Natrajan, R., Tan, D., Dowsett, M., Ashworth, A., and Reis-Filho, J. (2010c). Reflection and reaction, author' reply breast cancer molecular profiling with single sample predictors: a retrospective analysis. *The Lancet Oncology*, **11**(8), 720–721.

Weigelt, B., Baehner, F. L., and Reis-Filho, J. S. (2010d). The contribution of gene expression profiling to breast cancer classification, prognostication and prediction: a retrospective of the last decade. *Journal of Pathology*, **220**(2), 263–280.

Weinberg, R. A. (2007). *The biology of cancer*, volume 255. Garland Science New York.

Weng, L., Dai, H., Zhan, Y., He, Y., Stepaniants, S., and Bassett, D. (2006). Rosetta error model for gene expression analysis. *Bioinformatics*, **22**(9), 1111–1121.

Wessels, L., Reinders, M., Hart, A., Veenman, C., Dai, H., He, Y., and Veer, L. (2005). A protocol for building and evaluating predictors of disease state based on microarray data. *Bioinformatics*, **21**(19), 3755–3762.

Wirapati, P., Sotiriou, C., Kunkel, S., Farmer, P., Pradervand, S., Haibe-Kains, B., Desmedt, C., *et al.* (2008). Meta-analysis of gene expression profiles in breast cancer: toward a unified understanding of breast cancer subtyping and prognosis signatures. *Breast Cancer Research*, **10**(4), R65.

Wu, Z., Irizarry, R. A., Gentleman, R., Murillo, F. M., and Spencer, F. (2004). A model based background adjustment for oligonucleotide expression arrays.

Yasrebi, H., Sperisen, P., Praz, V., and Bucher, P. (2009). Can survival prediction be improved by merging gene expression data sets. *PloS ONE*, **4**(10), e7431.

Yu, J., Sieuwerts, A., Zhang, Y., Martens, J., Smid, M., Klijn, J., Wang, Y., and Foekens, J. (2007). Pathway analysis of gene signatures predicting metastasis of node-negative primary breast cancer. *BMC cancer*, **7**(1), 182.

Zakharkin, S., Kim, K., Mehta, T., Chen, L., Barnes, S., Scheirer, K., Parrish, R., Allison, D., and Page, G. (2005). Sources of variation in Affymetrix microarray experiments. *BMC Bioinformatics*, **6**(214), 1471–2105.

Zheng, W., Chung, L. M., and Zhao, H. (2011). Bias detection and correction in rna-sequencing data. *BMC bioinformatics*, **12**(1), 290.

Zhu, J., McLachlan, G., Ben-Tovim Jones, L., and Wood, I. (2008). On selection biases with prediction rules formed from gene expression data. *Journal of Statistical Planning and Inference*, **138**(2), 374–386.